Statistical–Dynamical Approach for Streamflow Modeling at Malakal, Sudan, on the White Nile River

Paul Block¹ and Balaji Rajagopalan²

Abstract: The upper White Nile Basin above Malakal, Sudan, is considered to be one of the most complicated and diverse hydrologic settings on Earth. Accurately depicting and predicting the streamflow at Malakal is essential for water managers considering Nile Basin-wide initiatives and potential large-scale projects. Dynamical, statistical, and combination models are assessed for their ability to predict monthly streamflow at Malakal. The dynamical model represents a lumped parameter, average-monthly water balance, whereas the statistical model incorporates a nonparametric approach based on local polynomial regression, utilizing principal components of precipitation and temperature. The combination of dynamical and statistical models through linear regression produces model weights of 0.44 and 0.59, respectively, implying a relatively balanced influence. Evaluation of the combination model demonstrates significant overall skill (correlation coefficients equal to 0.83), outperforming either individual model for the validation periods selected. Peak streamflow analyses of timing and quantity also exhibit superior performance by the combination model. An ensemble approach, practical for planning and management from a probabilistic standpoint, is additionally demonstrated.

DOI: 10.1061/(ASCE)1084-0699(2009)14:2(185)

CE Database subject headings: Dynamic models; Statistics; Streamflow; Sudan; Africa; Nile River.

Introduction

The upper White Nile Basin above Malakal, Sudan, is considered to be one of the most complicated and diverse hydrologic settings on Earth. The White Nile River originates in headstreams above Lake Victoria and continues nearly 2,900 km to Malakal, traversing through other lakes and the swamps of the Sudd in southern Sudan, draining a total of nearly 1.5 million km² (Shahin 1985). Fig. 1 illustrates the upper White Nile Basin, including the three equatorial lakes (Victoria, Kyoga, and Albert) and three swamp regions (Bahr el Jebel, Bahr el Ghazal, and Sobat) of primary interest; Malakal is located along the northern boundary of the basin. Both the lakes and the swamps in this basin exhibit nonlinear and discrete behavior. Additionally, the swamps impose a regulating effect, expanding during times of high inflow, allowing for increased levels of evaporation, and releasing at a more moderate rate unique to each swamp (Sutcliffe and Parks 1987). Alan Moorehead, in The White Nile, writing about the Sudd, said “there is no more formidable swamp in the world” (Moorehead 1971). Precipitation and evapotranspiration, along with the nonlinearities inherent in the system, are the driving forces behind the region’s hydrologic response.

Understanding and modeling streamflow within the Nile Basin is critical for effective water resources management, yet not trivial due to its size and composition, including ten riparian countries that share its water. To this end, accurate White Nile streamflow simulations and forecasts are highly desired. Yates and Strzepek (1998) created a dynamical model of the White Nile system, driven by common hydrometeorological inputs, capable of generating monthly flow scenarios under varying climatic conditions. The model is predominantly physically based, yet rather complex and fairly inflexible. To improve upon the dynamical model forecast capabilities, it is of interest whether combination with another model may bolster predictions. Recent developments in multimodel combination (Krishnamurti et al. 1999, 2000; Hagedorn et al. 2005; Rajagopalan et al. 2002; Regonda et al. 2006) suggest that combining outputs from different models tends to perform better than any single model. A statistically based modeling framework is proposed in this work for combination with the dynamical model, due to its simplicity in nature, with the hope of alleviating drawbacks associated with dynamical models, and capturing alternative streamflow features. The need for improved streamflow simulations is further motivated by a separate project focusing on hydropower implications along the Blue Nile in Ethiopia, with downstream ramifications in Sudan and Egypt (Block and Rajagopalan 2007).

This paper begins with a description of the data utilized, followed by background on contributing factors to Malakal streamflow variability. The three modeling frameworks, dynamical, statistical, and combination, are then presented. Model validation methods are subsequently described, followed by the results for Malakal streamflow assessment. The paper concludes with a summary and discussion of the results.

Data

Streamflow records at Malakal have been recorded monthly since 1912, and are publicly available until 1995. Numerous sources

---

¹Postdoctoral Research Scientist, International Research Institute for Climate and Society, Columbia Univ., Lamont Campus, 61 Rte. 9W, Palisades, NY 10964. E-mail: pblock@iri.columbia.edu
²Associate Professor, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado-Boulder, ECOT 441, 428 UCB, Boulder, CO 80309.
provide these data, including the National Center for Atmospheric Research’s (NCAR) ds552.1 dataset (Bodo 2001). In an attempt to diminish climate change influences, the data set is restricted to 1912–1990.

Input data for the dynamical and statistical models is derived from the Climate Research Unit’s (CRU) TS 2.0 and CL 2.0 data sets, obtained from the University of East Anglia (New et al. 2002; Mitchell et al. 2004). The first set includes monthly values for precipitation, mean daily temperature, diurnal temperature range, vapor pressure, and cloud cover; average monthly wind speeds were obtained from the second data set. Both sets are established on a 0.5° by 0.5° grid for 1901–2000, based on station data and anomalies. Although the historical station data within the region are sparse and spotty, the CRU precipitation data have been shown to be strongly correlated with other data sets, including the Climate Prediction Center’s merged analysis precipitation and the University of Delaware precipitation, for nearby regions (Block and Rajagopalan 2007).

Malakal Streamflow Variability

Monthly and annual time series for the recorded streamflow at Malakal, 1912–1990, are illustrated in Fig. 2. In Fig. 2(a), the solid line connects the monthly averages over the 79 year period, indicating high flow months in the later portion of the calendar year, specifically September through December. The boxes indicate the monthly variability in streamflow with the box covering the 25th and 75th percentile, the horizontal line within the box representing the median, and whiskers extending to the 5th and 95th percentiles. The dashed line presents average monthly precipitation over the same time period, clearly showing a distinct lead with streamflow, predominantly due to the regulating factor of the swamps in southern Sudan. From Fig. 2(b), it is evident that there have been two significantly high flow periods: the first between 1917 and 1918, and the second from 1963 to 1966. The 1964 event represents the highest annual flow on record. Both of these epochs are associated with periods of above normal precipitation in the upper White Nile region, as well as other neighboring equatorial regions (Block and Rajagopalan 2007). An additional factor coincident to the 1960s event was the revision of the water resources treaty between Uganda and Egypt for the relatively newly constructed Owens Falls Dam (1954) on Lake Victoria (Reynolds 2005).

The interannual and interdecadal variability in streamflow at Malakal and precipitation in the basin have been investigated by numerous researchers (Sutcliffe 1974; Kite 1981; WMO 1981; Shahin 1985; Sutcliffe and Parks 1987; Conway and Hulme 1993; Camberlin 1997; Mohamed et al. 2005). The summer rains near Malakal are part of the larger east African monsoon, resulting from a northward shift of the Intertropical Convergence Zone (ITCZ), a direct result of solar heating and warming of the surface (Griffiths 1972; Gamachu 1977). The main band of precipitation lies just south of the ITCZ. Malakal lies close to the northernmost extent of the ITCZ, and thereby receives the majority of its precipitation during July–September. Other parts of the basin southward of Malakal receive precipitation in a more even month-to-month distribution, or alternatively in two distinct seasons, due to the southward shifting of the ITCZ in the later calendar months of the year. Both the Atlantic and Indian Oceans act as contributing sources (Block and Rajagopalan 2007). Simultaneous to the shifting of the ITCZ, high-pressure systems in the South Atlantic and Indian Oceans, coupled with the Arabian and the Sudan thermal lows, allow for the influx of moisture into the basin (Seleshi and Zanke 2004).

Interannual variability is attributable to numerous climatic forcings, not the least of which is the El Niño southern oscillation phenomenon (Camberlin 1995; Nicholson and Kim 1997; Mutai and Ward 2000; Ntale and Gan 2004). It has also been suggested that precipitation in El Niño years is controlled by the Indian Ocean, whereas precipitation in La Niña years is directed by the Atlantic Ocean (Nicholson and Kim 1997). Other important factors include complex topography, the extent to which the ITCZ shifts, the equatorial lakes, and the Indian Ocean.

Modeling Framework

Numerous studies have developed dynamically based models of the Nile Basin, most recently under the pretense of studying potential climate change effects (Gleick 1991; Conway 1996, Strzepek et al. 1996; Yates and Strzepek 1998). Other models, such as the Nile Decision Support System (Georgakakos 2004), have been created in an effort to model basin hydrology and existing projects to assess basin-wide development scenarios. The goal of these types of models is essentially the same: to accurately depict or predict hydrologic conditions at various points of interest. Creating these models is not trivial, and often takes a significant amount of time for building and parameterization. The motivation of this study is to determine if a statistical model, much simpler in nature, may prove to be on par with existing...
dynamical models, and if a multimodel combination further enhances predictive capabilities.

For this work, two modeling frameworks are assessed, independently and in combination, for estimation of monthly Malakal streamflow, namely a dynamically based approach and a nonparametric local polynomial statistical approach. The nonparametric approach has been selected in lieu of traditional parametric techniques due to the complex, inherent nonlinearities within the basin. Fig. 3 clearly presents the nonlinear relationship between Malakal streamflow and basin-wide average precipitation and temperature, two dominant factors in determining streamflow quantity. It is the capability of the nonparametric technique to capture these nonlinearities that makes it so attractive.

Dynamical Model

The dynamical model utilized in this study is a modification of the full Nile Basin-wide model developed by Yates and Strzepek (1998) to assess streamflow variations under climatic change. It is a simple, lumped parameter, average-monthly water balance model incorporating six subbasins above Malakal, each associated with one of the equatorial lakes or swamps [see Fig. 2 in Yates and Strzepek (1998)]. Three major components constitute the model, including soil moisture accounting, evaluation of potential evapotranspiration, and a reservoir storage scheme for both lakes and swamps. Fig. 3 in Yates and Strzepek (1998) depicts the water balance component of the model.

Model inputs include precipitation, temperature, vapor pressure, cloud cover, and winds to produce monthly runoff and evapotranspiration aggregated within each subbasin. Direct precipitation and evaporation over the lakes and swamps is also included. The outlet works for each lake are unique, and based on nonlinear lake level–discharge relationships. The discharge from each swamp is a function of the swamp depth, recharge coefficient, lateral spread, and lagging of upstream inflows.

The model may be run for a user-defined number of years, with outputs including lake levels and flow rates at points within each subbasin. For the purposes of this study, only the streamflow rate at Malakal is reported.

Local Polynomial Statistical Model

Nonparametric statistical methods are becoming increasingly popular for modeling in hydroclimatological studies (Lall 1995; Regonda et al. 2005; Prairie et al. 2007; Grantz et al. 2007; Block and Rajagopalan 2007). These methods provide an attractive alternative for addressing the drawbacks of traditional linear regression, including meeting normality requirements, the potentially large influence by a small number of outliers, and the inability to capture nonlinear relationships between the dependent and independent variables. Generally, regression models take the following simple form:

\[ Y = f(x) + e \]  

(1)

where \( x \) represents a vector of regression variables (independent variables); \( f \) = function; \( Y \) = dependent variable and \( e \) = error, often assumed to be normally distributed with a mean of zero and variance \( \sigma^2_e \). Traditional linear regression involves fitting a linear function \( f \) to the entire data. In the nonparametric approach, estimation of the function \( f \) is performed “locally” at the point to be estimated. This local estimation provides the ability to capture...
features (i.e., nonlinearities) that might be present locally, without granting outliers any undue influence in the overall fit (Block and Rajagopalan 2007). Several nonparametric methods for regression and probability density function estimation exist; for an overview of these methods and their applications to hydroclimatology, see Lall (1995).

In this work, the local polynomial-based nonparametric approach (Loader 1999) is proposed for its ease in understanding, implementation, and successful past applications. The methodology is described in the following algorithm (Block and Rajagopalan 2007). For a point of interest where an estimation of the function is desired, say \( x_{pl} \) (subscript \( p \) represents “predictive,” and \( l \) represents “local”):

1. \( K (=\alpha N) \) nearest neighbors are identified in proximity to \( x_{pl} \). The neighbors can be obtained using either the Euclidean or Mahalanobis distance. The parameter \( \alpha \) describes the size of the neighborhood and is within the \((0, 1] \) range. \( N \) represents the total number of data points. Clearly, if \( \alpha \) takes a value of 1, the number of neighbors selected includes all data points.
2. A polynomial of order \( P \) is fit to the \( K \) nearest neighbors, using a weighted least-squares method. The fitted polynomial is used to obtain the estimate of the dependent variable, \( Y_{pl} \). The local error standard deviation, \( \sigma_{pl} \), can be obtained from regression theory (Helsel and Hirsch 1995).

It is noteworthy to mention that for \( \alpha = 1 \) and \( P = 1 \), this approach still differs from the traditional linear regression, due to the weighting scheme. Linear regression utilizes ordinary least squares for optimization of predictor coefficients, which effectively weights each point equally. The local polynomial approach uses a weighted least-squares scheme, weighting observed data points closer to \( x_{pl} \) higher. This gives more influence to local points, and little to none for points far from the desired prediction point. For a perfectly linear relationship, the results of the local polynomial approach are indistinguishable from the traditional linear regression approach results, thus making it a more general and flexible approach.

The advantages of this approach over the dynamical model are clear: the model is a great deal simpler to create and code, takes less processing time, requires significantly fewer inputs, and is not required to match intermediary flows in the system, only streamflow at Malakal.

### Generalized Cross-Validation Skill Score

The optimal values of the two parameters \( K \) (or \( \alpha \)) and \( P \) must be estimated from the historical data, and may be obtained using the generalized cross-validation (GCV) score function, given in the following:

\[
GCV(\alpha, P) = \frac{\sum_{i=1}^{N} e_i^2}{\frac{N}{1 - \frac{m}{N}}} 
\]

where \( e_i \) = model residual (difference between observed and model-estimated values of the dependent variable), and \( m \) = number of regression variables in the fitted polynomial. The GCV function penalizes overfitting (large numbers of regression
variables) and is a very good estimate of the predictive risk (Craven and Wahba 1979).

For each combination of $\alpha$ and $P$, the model is fitted, as described in the above presented algorithm, and the GCV score is computed; the combination providing the minimum GCV score is selected as the optimal one.

The GCV function can also be used to select the best subset from a suite of regression variables. This process involves including different combinations of the regression variables, along with varying $\alpha$ and $P$ values, calculating the GCV, and selecting the combination of regression variables, $\alpha$, and $P$ that provide the minimum GCV score as the best parameter combination. The use of GCV for subset selection is fairly recent (Regonda et al. 2005, 2006) and has been shown to be quite effective. The application of this method in the present research is described in the following.

**Identification of Best Variables for Local Polynomial Model**

Potential regression variables for the local polynomial statistical model were limited to inputs utilized in the dynamical model, including precipitation, temperature, and potential evapotranspiration. The latter was eventually eliminated, as it is strongly correlated to both precipitation and temperature. To capture the variability of the entire upper White Nile Basin, six representative precipitation and temperature locations were chosen throughout, mimicking the dynamical model subbasins, as illustrated by the triangles in Fig. 1, totaling twelve variables. Not surprisingly, the variables demonstrated multicollinearities (i.e., strongly correlated amongst each other), even though the spatial region considered is quite large. To combat this, principal component analyses (PCA) were performed separately on the precipitation and temperature data. PCA methods, widely used in climate research, decompose a space–time random field (e.g., multivariate data set such as the monthly precipitation at the six locations in the basin) into orthogonal space and time patterns using eigendecomposition (von Storch and Zwiers 1999). The space–time patterns (also called “modes”) are ordered according to the percentage of variance captured. Typically, the first few modes capture most of the variance present in the data. This is analogous to a dimension reduction technique, where a large multivariate data set is effectively represented by a few modes (i.e., smaller dimension). The temporal patterns are also referred to as principal components (PCs). As the PCs are orthogonal, they can be analyzed independently and combined to reconstruct the original data.

The mathematical formulation is as follows:

$$\text{PC} = [E][X]$$

where $X=$ data matrix; $E=$ matrix of eigenvectors; and $\text{PC}=$ corresponding matrix of the principal components.

The variance explained by each mode, also known as the eigenspectrum, from the PCA of precipitation and temperature for the 1912–1990 period is shown in Fig. 4. Clearly the leading two modes capture most of the data variance. Thus, the twelve PCs, six each from precipitation and temperature, form the potential variable set used in the above-described GCV framework for selecting the best subset.

**Combination Model**

A combination model, or multimodel, approach is also undertaken for evaluation of its potential ability to capture Malakal streamflow. Previous researchers have indicated that multimodel techniques may produce more robust results than single model approaches (Morel-Seytoux et al. 1993; Balmaseda et al. 1994; Regonda et al. 2006). Additionally, the National Weather Service

![Fig. 4. Eigenspectrums from principal component analyses for (a) precipitation; (b) temperature](image)
currently uses a statistical–dynamical multimodel approach for long-lead forecasts of atmospheric variables (Ed O’lenic, personal communication, May 23, 2006). For this work, we propose the optimal combination of dynamical and statistical models using a simple linear regression, with an assumed $y$ intercept of zero, over the entire 1912–1990 period. The choice of linear regression is based on the ability to easily assess the relative influence of each model and gauge its general contribution. Eq. (4) presents the following relationship:

$$SF_M = \beta_1 SF_D + \beta_2 SF_S + e$$

where $SF_D$=streamflow at Malakal as determined by the dynamical model; $SF_S$=streamflow as determined by the statistical model; $e$=model error; $\beta_1$ and $\beta_2$=optimal coefficients; and $SF_M$=combination model streamflow estimation.

### Model Calibration and Validation

The dynamical model was calibrated using average monthly parameters for each input variable over 1948–1973 by minimizing errors between mean modeled streamflow and mean observed streamflow (Yates and Strzepek 1998). The authors claim that this calibration aptly captures the hydrologic features of the basin. All validation periods presented in this study are based on the aforementioned calibration period. Further details are available in Yates and Strzepek (1998).

The statistical model is validated through a cross-validation approach by dropping a fraction of the data (i.e., the validation period), obtaining the best subset of model variables, fitting a model on the remaining data (i.e., the fitting period), and then predicting the monthly streamflow for the validation period. The

<table>
<thead>
<tr>
<th>Period</th>
<th>Dynamical</th>
<th>Statistical</th>
<th>Combination</th>
<th>Statistical predicted</th>
<th>Combination predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1912–1990</td>
<td>0.75</td>
<td>0.77</td>
<td>0.83</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1912–1950, 1971–1990 (C)</td>
<td>0.70</td>
<td>0.79</td>
<td>0.82</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1951–1970 (V)</td>
<td>0.83</td>
<td>—</td>
<td>—</td>
<td>0.70</td>
<td>0.83</td>
</tr>
<tr>
<td>1912–1970 (C)</td>
<td>0.79</td>
<td>0.75</td>
<td>0.83</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1971–1990 (V)</td>
<td>0.64</td>
<td>—</td>
<td>—</td>
<td>0.86</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Note: C=calibration period; and V=validation period.

![Fig. 5. Observed (solid line) and combination model (dashed line) streamflow at Malakal, 1951–1970](image)
necessary PCs for the validation period are obtained by multiplying the variable (precipitation or temperature) by the eigenvalues from the PCA of the fitting period.

To test the statistical model robustly, validation was performed in three different ways: (1) dropping each year individually (mimicking a prediction-type assessment); (2) validating on 1951–1970; and (3) validating on 1971–1990. The period 1951–1970 represents a sharp increase followed by a decline in Malakal streamflow, including the wettest year on record, whereas the 1971–1990 period portrays a steady decline. An examination of annual streamflow in Fig. 2 visually illustrates these patterns. The GCV approach described earlier was utilized to obtain the optimal model parameters $\alpha$ and $P$, and the best subset of variables for the fitting periods. Using GCV, the best subset of variables includes the first two PCs of both precipitation and temperature as well as the first PC for each lagged by 1 month. These six predictors constitute the optimal set for all validation periods. Inclusion of the lagged PCs as a predictor is not surprising, due to the regulating characteristics of the swamps in the northern half of the study area. The best order of polynomial $P$ was found to be 1 for all periods and $\alpha$ was found to be 0.20 when dropping each year individually, 0.25 for 1951–1970, and 0.3 for 1971–1990.

The predictions resulting from the dynamical, statistical, and combination models are evaluated by correlation coefficients between the model predictions and the observed streamflow values.

### Results

Correlation coefficients between streamflow predictions from the dynamical, statistical, and combination models and historical observations are presented in Table 1. Over the entire 1912–1990 period, the dynamical and statistical models are quite comparable, with correlation coefficients of 0.75 and 0.77, respectively. Not surprisingly, the combination model produces a higher correlation coefficient of 0.83, by taking advantage of the capabilities of both models. The combination model coefficients, $\beta_1$ and $\beta_2$, as presented in Eq. (4), are 0.44 and 0.59 for the dynamical and statistical models, respectively, implying a relatively balanced influence, with slightly greater weight originating from the statistical model.

The two 20 year validation periods also prove insightful. The combination model produces a strong correlation coefficient of 0.83 for both validation periods, giving credence to the ability of the model to perform robustly throughout varying climatic trends. This is not the case, however, for the dynamical and statistical models independently. During the 1951–1970 validation period, the correlation coefficient for the dynamical model is quite high (0.83), whereas the statistical model is less (0.70). The 1971–1990 validation period, however, illustrates a reversal, as the statistical model produces a high correlation coefficient (0.86), with the dynamical model significantly lower (0.64). These two epochs
clearly demonstrate the advantages of the combination approach, as it tempers the inadequacies of the individual models, but is able to capture the positive features of both. Not unexpectedly, the optimal combination model coefficients utilized in the two validation periods also reflect the ability of the dynamical and statistical models to capture streamflow during the associated calibration periods. For 1951–1970, the combination model coefficients come to 0.55 and 0.5, favoring the dynamical model; coefficients for 1971–1990 are 0.34 and 0.69, favoring the statistical model.

Figs. 5 and 6 illustrate the observed and combination model predicted monthly streamflow at Malakal over the two 20 year validation periods. The combination model results mimic the observed values quite well, with the exception of grossly underestimating the 1964 peak and overestimating the 1987 peak. Figs. 7 and 8 show fewer years, 1962–1966, including the wettest year in the record, and 1988–1990, respectively, with the dynamical and statistical model results also included. The 1988–1990 figure is enlightening, illustrating the ability of the combination model to closely approximate the peak in the first 2 years, even though the two independent models deviate significantly; it fails, however, to adequately capture the 1990 peak.

A peak streamflow analysis reveals the ability of the dynamical, statistical, and combination models to predict both the timing and quantity of the highest monthly streamflow in each year. Fig. 9 illustrates the capacity of the models to predict the month in which the peak streamflow occurs. Clearly, the statistical model histogram is more tightly grouped around the highest observed month than the dynamical model histogram, thereby assisting the combination model in better forecasting the peak timing. Virtually all (97%) of the combination model predictions of peak streamflow lie within 1 month of the observed peak month, with a strong majority falling into the correct month. Fig. 10 indicates each model’s capacity to capture the peak streamflow quantity. Although the dynamical model tends to better reproduce higher peaks, the statistical model (and therefore the combination model) generally appears to undersimulate peak streamflow. This is a direct result of the statistical modeling nature and limited data availability; when the statistical model is forecasting a wet year, and the fitting period includes few significant wet years, prediction will prove difficult. Correlation coefficient skill scores for the dynamical, statistical, and combination models equate to 0.47, 0.2, and 0.47, respectively, favoring the dynamical model, whereas root mean square errors total to 7,067, 6,312, and 5,673 for peak streamflow quantity, favoring the statistical model. Overall, the combination model is clearly superior, and appears to alleviate timing and quantity deficiencies evident in the two individual models.

An ensemble approach, in lieu of the deterministic results described thus far, may also be advantageous. This process involves use of the statistical model errors to create random normal deviates, with mean zero and variance $\sigma^2_{\text{stat}}$, which are added to the predicted streamflow values from the statistical model, thus providing ensembles and the associated probability density function. This has been used successfully in predicting seasonal rainfall and streamflow (see, e.g., Grantz et al. 2007; Singhrattna et al. 2005; Regonda et al. 2006). The combination model ensemble is generated using Eq. (4), with monthly stochastic predictions from the statistical model. Combination model coefficients remain unchanged. Fig. 11 illustrates the combination model ensemble for the 1988–1990 period. Ensembles are shown as box
plots in Fig. 11, with the box covering the 25th and 75th percentile, the horizontal line inside the box representing the median, whiskers extending to the 5th and 95th percentile, and outliers shown as circles.

The rank probability skill score (RPSS), a measure of the skill of ensemble forecasts, is a widely used probabilistic measure for comparison with prediction by climatology forecasts (Wilks 1995; Saunders and Fletcher 2004). The general form of the rank probability score (RPS) equation for any year takes the form:

$$RPS = \frac{R}{\sum_{m=1}^{R} (CP_{F,m} - CP_{O,m})^2}$$  \hspace{5em} (5)

where $R$=number of categories; $CP_{F,m}$=cumulative predicted probability for the forecast ensemble (through category $m$); and $CP_{O,m}$=cumulative observed probability (also through category $m$). This study incorporates three categories of equal size (e.g., below normal, near normal, or above normal streamflow), such that the climatological probability of being in each category is $1/3$; for the category that was observed the probability is one, and zero elsewhere. A perfect forecast results in RPS equaling zero. The RPSS is then defined as

$$RPSS = 1 - \frac{RPS_{\text{forecast}}}{RPS_{\text{climatology}}}$$  \hspace{5em} (6)

RPSS values range from +1 to $-\infty$. A value of +1 represents perfect skill, or a perfect forecast, whereas negative values represent poor skill; any value above zero represents an improved forecast over climatology. The RPSS is calculated for each year separately. For the ensemble forecast shown in Fig. 11, a median RPSS skill score of 0.65 results. This indicates a skillful ensemble forecast and gives clear indication of the improvement over climatology. The ensemble approach provides a framework for probability density function assessment and evaluation of threshold exceedance probabilities, especially useful for probabilistic prediction of flood or drought conditions (e.g., Block and Rajagopalan 2007; Grantz et al. 2007).

**Summary and Discussion**

Dynamical, statistical, and combination models are assessed for their ability to predict monthly streamflow at Malakal, Sudan. The statistical model incorporates a nonparametric approach based on local polynomial regression, utilizing principal components of...
Fig. 9. Peak flow timing histogram for (a) dynamical; (b) statistical; and (c) combination models compared to observed flows over 1912–1990. Zero indicates modeled peak flow months are identical to observed peak flow months, ±1 indicates model prediction differs by 1 month, etc.

Fig. 10. Peak flow quantity histogram for (a) dynamical; (b) statistical; and (c) combination models compared to observed flows over 1912–1990. Diagonal line represents perfect predictions.
precipitation and temperature throughout the upper White Nile Basin. The GCV function is employed for determining the best subset (six) of regression variables out of a suite of twelve potential ones. The combination model is a simple linear regression of the outputs from the dynamical and statistical models; evaluation of correlation coefficients for the combination model demonstrates significant overall skill, outperforming either of the other two models independently. Peak streamflow analyses of timing and quantity also exhibit superior performance by the combination model. An ensemble approach to the combination model, using random normal deviates of the statistical model error, is also illustrated for 1988–1990, and demonstrates a framework for planning and management from a probabilistic standpoint.

The combination model could easily be transformed into a prediction tool if forecasts of precipitation and temperature are available, making it attractive to basin managers and decision makers. Other options include implementing precipitation and temperature scenarios from general circulation models for assessing potential climate change effects.

Other aspects also warrant further attention, including threshold exceedance probability evaluation and extending the model to incorporate a larger portion of the Nile Basin. By applying the ensemble forecasting framework to the model, threshold exceedances may be set to give indication of risk levels of wet or dry (or flood or drought) conditions in a month or season (Block and Rajagopalan 2007). Extending the model, particularly to the Aswan Dam in Egypt, would also be of value for comparison with existing basin-wide water systems models and evaluation of potential large-scale basin projects.

Acknowledgments

This research study was partially funded by a grant from the U.S. Agency for International Development (USAID) through the International Food Policy Research Institute (IFPRI), and forms part of the first writer’s Ph.D. dissertation at the University of Colorado—Boulder. The writers wish to express their appreciation to the editors and three anonymous reviewers for their insightful comments and suggestions, undeniably improving the quality of this paper.

References


Fig. 11. Box plots of monthly ensemble predictions for 1988–1990, including observed flows (solid line) and mean combination model predictions (dashed line)