Journal of Hydrology 452–453 (2012) 226–231

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A toy model for monthly river flow forecasting

Xubin Zeng [a,*], Kira L. Kiviat [a], Koichi Sakaguchi [a], Alaa M.A. Mahmoud [b]

[a] Department of Atmospheric Sciences, The University of Arizona, Tucson, AZ, United States
[b] Department of Astronomy and Meteorology, Al-Azhar University, Faculty of Science, Nasr City, Cairo 11884, Egypt

ARTICLE INFO

Article history:
Received 12 March 2012
Received in revised form 15 May 2012
Accepted 26 May 2012
Available online 4 June 2012
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Ashish Sharma, Associate Editor

Keywords:
River flow forecasting
Neural network
Uncertainty quantification
Modified Nash–Sutcliffe coefficient of efficiency

SUMMARY

River flow forecasting depends on land–atmosphere coupled processes, and is relevant to hydrological applications and land–ocean coupling. A toy model is developed here for monthly river flow forecasting using the river flow and river basin averaged precipitation in prior month. Model coefficients are calibrated for each month using historical data. The toy model is based on water balance, easy to use and reproduce, and robust to calibrate with a short period of data. For five major rivers in the world, its results agree with observations very well. Its prediction uncertainty can be quantified using the model's error statistics or using a dynamic approach, but not by the dispersion of 10,000 ensemble members with different sets of coefficients in the model. Its results are much better than those from a physically based land model even after the mean bias correction. The toy model and a standard neural network available from the MATLAB give similar results, but the latter is more sensitive to the length of calibration period. For the monthly prediction of river flow with a strong seasonal cycle, a modified Nash–Sutcliffe coefficient of efficiency is introduced and is found to be more reliable in model evaluations than the original coefficient of efficiency or the correlation coefficient.

1. Introduction

River flow forecasting is one of the essential issues in applied hydrology; e.g., for water resources management and flood control (Nash and Sutcliffe, 1970). Discharges from major rivers also provide significant freshwater at the mouth of these rivers to the oceans. These discharges balance the freshwater loss due to ocean surface evaporation and affect regional ocean circulations through changes in density. Furthermore, they bring large amounts of particulate and dissolved minerals and nutrients to the oceans and hence affect the global biogeochemical cycles (Dai et al., 2009).

In general, river flow forecasting depends on land–atmosphere interactions. Precipitation and snowmelt represent the natural source of water in the river, while evapotranspiration is the sink. Human activities (e.g., irrigation; flow regulation due to dam and reservoir operations) represent additional source and sink. River flow is also affected by the three-dimensional movement of soil water which is dependent on many factors such as the topography, soil texture, vegetation cover, and groundwater aquifers. Physically-based land surface models or hydrological models along with river routing schemes attempt to explicitly simulate most of these processes (Lohmann et al., 2004). This approach has also been used recently in the ensemble streamflow prediction system developed by the US National Weather Service for water resources outlook (including the monthly and seasonal streamflow outlook) (Noel et al., 2010).

Recognizing the lack of comprehensive data that are required for physically-based models, a variety of data-driven models have also been used for river flow forecasting, among which an artificial neural network (Karunanithi et al., 1994; Hsu et al., 1995) is the most widely used. However, the network structure, including the number of hidden nodes, is difficult to determine in general and has to be developed through a trial-and-error approach.

While progress has been continuously made in both land modeling and neural network, here we take a very different approach by combining the strengths of land modeling (i.e., physically-based) and neural network (i.e., data-driven) to develop a simple nonlinear model. This physically-based and data-driven model contains a single equation for the river flow, and hence is referred to as a toy model. The goal of this paper is to develop such a toy model and compare it with a land surface model and a neural network using observations over five major rivers (see Table 1) in the world.

2. Model descriptions

2.1. The toy model

Based on water balance, the river flow F (km^3 or billion cubic meters) can be predicted from

* Corresponding author. Address: Department of Atmospheric Sciences, 1118 E. 4th St., The University of Arizona, Tucson, AZ 85721, United States. Tel.: +1 520 621 4782; fax: +1 520 621 6833.
E-mail address: xubin@atmo.arizona.edu (X. Zeng).

0022-1694/$ - see front matter
http://dx.doi.org/10.1016/j.jhydrol.2012.05.053

Table 1
Station information. Lat/Lon: latitude/longitude; Cal/Val: calibration/validation; F_{ave} (km^3/month) and P_{ave} (mm/month): average river flow and precipitation for the whole period.

River	Station	Lat/Lon	Cal/Val periods	F_{ave}	P_{ave}
Mississippi	Vicksburg	32.3 °N/90.9 °W	1932–1981/1982–1998	44.89	89.18
Nile	El-Ekhsase	29.7 °N/31.3 °E	1948–1987/1988–2003	1.45	0.53
Yangtze	Datong	30.8 °N/117.6 °E	1901–1943/1946–2000	76.26	57.65
Amazon	Obidos	1.95 °S/55.5 °W	1928–1977/1978–2006	450.81	450.81
Colorado	Below Hoover Dam	36.0 °N/114.7 °W	1935–1984/1985–2005	1.04	8.82

$$\frac{dF}{dt} = \alpha(P - c) - \gamma F \quad (1)$$

where P is the observed monthly precipitation (including rainfall and snowfall) over the river basin (mm/month), and c includes evapotranspiration and snowfall (represented by positive values) as well as snowmelt (represented by negative values) (mm/month). Instead of explicitly computing c, we determine it for each month through calibration. The coefficient α provides the conversion between the water input over a river basin and the river flow. It also includes human activities (such as irrigation and flow regulation) that are not explicitly modeled. The first term on the right-hand side represents the net water input to river flow. The last term represents the river water level recession. If γ is a constant, $1/\gamma$ would represent the e-folding time of the recession. In general, γ does not need to be a constant and is assumed to be

$$\gamma = \beta(F/F_{ave})^d \quad (2)$$

where F_{ave} refers to the climatological average of F (see Table 1), and coefficient β ($month^{-1}$) and exponent d are determined for each month.

Eqs. (1) and (2) can be discretized as

$$\frac{F_{n+1} - F_n}{\Delta t} = \alpha_n(P_n - c_n) - \beta_n(F_n/F_{ave})^{d_n} F_{n+1} \quad (3)$$

where the subscript n refers to month and Δt refers to the time step of 1 month. Assuming $a_n = \alpha_n \Delta t$, and $b_n = \beta_n \Delta t$, Eq. (3) can be rearranged as

$$F_{n+1} = \frac{a_n(P_n - c_n) + F_n}{1 + b_n(F_n/F_{ave})^{d_n}} \quad (4)$$

For the river flow at a particular station, variables a_n, b_n, c_n, and d_n in Eq. (4) are calibrated for each month using historical flow data. These values are independent of years, and their constraints include: a_n and b_n are positive, c_n can be positive (evapotranspiration or snowfall) or negative (snowmelt), and d_n is between 0 and 1.

The range and increment for a_n and c_n differ across rivers according to the climatological flow and precipitation characteristics, such as the average, minimum, and maximum values. For example, the calibration for the Mississippi River used values from 0.15 to 1.62 (km^3 month/mm) for a_n (divided into 100 evenly spaced values), and from −83.0 to 218.5 (mm/month) (divided into 70 evenly spaced values) for c_n, while for the Amazon River a_n varies from 0.23 to 2.58 (km^3 month/mm) and c_n varies from −335.9 to 884.0 (mm/month). The range and increment of b_n or d_n are the same across stations (b_n: from 0 to 2.98 with 0.02 increment; d_n: from 0 to 1 with 0.1 increment). Calibration is done simply by evaluating the monthly forecast using Eq. (4) with every combination of assignments to the variables (out of about 1.2×10^7 total combinations). The chosen values for each month would minimize the mean square error for that month during the calibration period. Results are insensitive to the exact increment for each variable (not shown).

2.2. Neural network

There are numerous neural network models used for river flow forecasting (e.g., El-Shafie and Noureldin, 2011; Nilsson et al., 2006). Here we use the standard multilayer perceptron (MLP) neural network, which is the most commonly used and consists of multiple layers: an input layer, output layer, and one or more "hidden" layers. The software is available from the MATLAB Neural Network Toolbox (http://www.mathworks.com/products/neural-net/).

The input layer includes multiple variables (or nodes) such as monthly river flow and precipitation as described below. Only one hidden layer with multiple nodes is used here with each node depending on the nodes in the input layer using different transfer functions. The output layer contains only one node which is computed using nodes in the hidden layer through different transfer functions.

The inputs to the network can be determined through cross-correlation analysis of the historical flow data with itself and with other data (El-Shafie and Noureldin, 2011). For our cases, the chosen input values (for predicting river flow at month $n + 1$) include river flow at months n and $n - 1$, and precipitation at month n. Furthermore, because there is a significant seasonal variation of river flow, pairs of values from the sine and cosine functions of the seasonal cycle (Nilsson et al., 2006) are used as input variables as well. The calibration and validation periods are the same as those for the toy model.

The optimal number of nodes in the hidden layer and the transfer function are determined through trial and error using various algorithms available from the MATLAB. For our cases, log sigmoid (for the Yangtze River) and tangent sigmoid (for other four rivers) transfer functions are used. The number of hidden nodes also varies from 5 for the Yangtze and 6 for the Mississippi to 10 for other three rivers. The Bayesian Regularization algorithm is used for training.

2.3. Community Land Model (CLM4)

The Community Land Model (CLM4) (Lawrence et al., 2011) is the land component used in the Community Earth System Model. Biogeophysical processes simulated by CLM4 include solar and longwave radiation interactions with vegetation canopy and soil, turbulent fluxes from canopy and soil, heat transfer in soil and snow, hydrology of canopy, soil, and snow, and stomatal physiology and photosynthesis.

For the offline CLM4 simulations, 3-hourly atmospheric forcing data of downward solar radiation, downward longwave radiation, rainfall and snowfall, wind, temperature, and humidity are used to produce surface turbulent and radiative fluxes for coupling to the atmosphere, to produce vertical distribution of soil temperature and moisture for coupling to ecosystem dynamics, and to produce surface and subsurface runoff for coupling to the oceans. Specifically, soil water removed from the model runoff is moved horizontally in a simple routing scheme (Lawrence et al., 2011).