The US MOPEX Data Set

JOHN SCHAAKE[1], SHUZHENG CONG[1] & QINGYUN DUAN[2]

1 *Office of Hydrologic Development, NOAA National Weather Service, 1325 East–West Avenue, Silver Spring, Maryland 20910, USA*
john.schaake@noaa.gov

2 *Lawrence Livermore National Laboratory, Energy and Environment Directorate, 7000 East Avenue, Livermore, California 94550, USA*

Abstract A key step in applying land surface parameterization schemes is to estimate model parameters that vary spatially and are unique to each computational element. Improved methods for parameter estimation (especially for parameters important to runoff response) are needed and require data from a wide range of climate regimes throughout the world. Accordingly, the GEWEX Hydrometeorology Panel (GHP) endorsed the concept of an international Model Parameter Estimation Project (MOPEX) at its Toronto meeting, August 1996. Phase I of MOPEX was funded by NOAA in financial year (FY) 1997, Phase II in FY 2000 and Phase III in FY 2003. MOPEX was adopted as a project of the IAHS/WMO Working Group on GEWEX and of the WMO Commission on Hydrology (CHy), and is now a contributor to the Combined Enhanced Observing Period (CEOP) of the World Climate Research Program (WCRP). In 2004, MOPEX became a Working Group of the IAHS Prediction in Ungauged Basins (PUB) Initiative. MOPEX is also expected to contribute to the work of the Hydrologic Ensemble Prediction Experiment (HEPEX). The primary goal of MOPEX is to develop techniques for the *a priori* estimation of the parameters used in land surface parameterization schemes of atmospheric models and in hydrological models. A major early effort of MOPEX has been to assemble a large number of high-quality historical hydrometeorological and river basin characteristics data sets for a wide range of river basins (500–10 000 km^2) throughout the world. MOPEX data sets are available via the Internet (ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/). This paper documents the development of data sets for river basins in the USA. Several highly successful parameter estimation workshops have been organized by MOPEX. The first was held as part of the IAHS General Assembly in Birmingham, UK, in July 1999. The second workshop was hosted in April 2002 in Tucson, Arizona, USA, by the SAHRA/University of Arizona. The third MOPEX workshop was held as part of the IAHS General Assembly in Sapporo, Japan, July 2003. The fourth, in Paris, France, July 2005 was organized by the CEMAGREF in collaboration with the ENGREF, Météo France, National Weather Service and the SAHRA/University of Arizona. The fifth workshop was held as part of the IAHS Scientific Assembly, February 2005, Foz do Iguacu, Brazil. The purpose of the future phases of the project is to: (a) continue collecting additional international data sets; update data from the USA by adding recent years, including data for elevation zones in mountainous areas and refining energy forcing data; (b) continue to conduct international MOPEX workshops; (c) provide leadership to develop a better scientific understanding of how to improve procedures for *a priori* parameter estimation; (d) make a significant hydrological contribution to CEOP and PUB; and (e) demonstrate transferability of MOPEX results. The basic data collection strategy being used in MOPEX is to seek the most readily available and highest quality data first. During the next three years, analyses of the available MOPEX data sets by the international scientific community will be emphasized.

Key words *a priori*; calibration; hydrological models; parameters; regionalization; transferability; uncertainty

INTRODUCTION

Background

A critical step in applying a hydrological model to a basin or a land surface parameterization scheme (LSPS) of an atmospheric model to a specific grid element is to estimate the coefficients or constants (i.e. parameters) in the model. Parameters are inherent in all models. In general, they vary spatially so they are unique to each basin or grid point. Some may also vary seasonally. Moreover, some parameters may be space-time scale dependent.

A common approach in the hydrological modelling community to parameter estimation is to calibrate hydrological models to historical observations by tuning model parameters. For ungauged basins and for LSPS applications, it is difficult to obtain adequate data needed for model calibration. A further complication is that LSPSs are typically applied to large spatial scales and involve many grid elements. To estimate model parameters in those cases, it is necessary to assign model parameters *a priori*.

A priori parameter estimation procedures are available for many hydrological models and LSPSs. But these procedures have not been fully validated through rigorous testing using retrospective hydrometeorological data and corresponding land surface characteristics data. This is partly because the necessary database needed for such testing has not been available until recently. Moreover, there is a gap in our understanding of the links between model parameters and the land surface characteristics. Generally available information about soils (e.g. texture) and vegetation (e.g. type or vegetation index) only indirectly relates to model parameters such as the hydraulic properties of soils and rooting depths of vegetation. Also, it is not clear how heterogeneity associated with spatial land surface characteristics data affects those characteristics at the scale of a basin or a grid cell. Consequently, there is a considerable degree of uncertainty associated with the parameters given by existing *a priori* procedures. It is necessary to develop enhanced *a priori* parameter estimation methodologies for hydrological models and LSPSs. Toward this goal, a project known as the Model Parameter Estimation Experiment (MOPEX) was initiated in 1996. The MOPEX project has been truly an international collaborative endeavour, with the involvement of international scientists and hydrological data assembled from different countries.

MOPEX Science Strategy

The MOPEX Science Strategy involves three major steps, as illustrated in Fig. 1. The first step is to develop the necessary data sets. The next is to use these data to develop *a priori* parameter estimation methodology. Step three is to demonstrate that new *a priori* techniques produce better model results than existing *a priori* techniques for basins not used to develop the new *a priori* techniques.

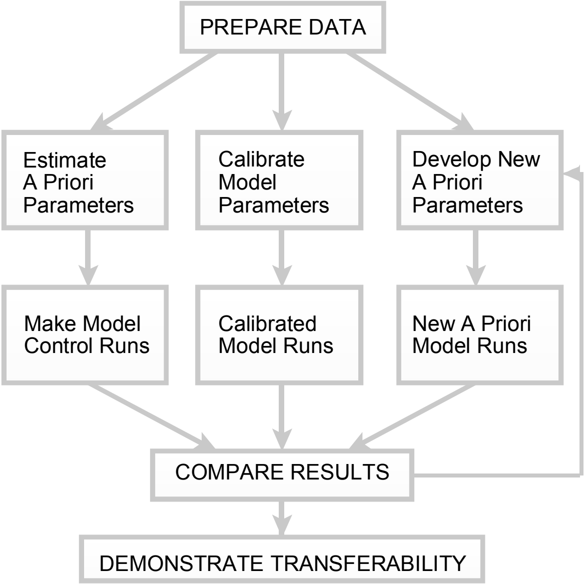
PREPARE DATA
Estimate A Priori Parameters
Calibrate Model Parameters
Develop New A Priori Parameters
Make Model Control Runs
Calibrated Model Runs
New A Priori Model Runs
COMPARE RESULTS
DEMONSTRATE TRANSFERABILITY

Fig. 1 The MOPEX Science Strategy.

Step two is accomplished using a three-path strategy. The first path is to make reference runs with model parameters estimated by using existing *a priori* parameter estimation procedures. The second path is to make model runs using calibrated or tuned values of selected model parameters. Then, the calibrated parameters are analysed to improve the relationships between model parameters and basin characteristics including climate, soils, vegetation and topographic features. The new relationships are then used to estimate the new *a priori* parameters. The third path is to make new model runs using the new *a priori* parameter estimates. The success of step two is measured by how much improvement in model performance is achieved when the model is operated using new *a priori* parameters as compared to the reference runs.

DATA REQUIREMENTS

The data required for MOPEX can be grouped into four categories:

(a) basic required observations for development and testing;
(b) required physical characteristics;
(c) desirable additional observations;
(d) observations for detailed testing and evaluation.

These are explained below. In each case, the minimum data required are believed to be readily available for the basins to be used. The desired level of data are also believed to be available in many basins.

Basic required observations for development and testing

Historical/retrospective data are needed for many years (as long as possible, e.g. in the USA the period 1948–to date) for at least several hundred test basins which have the minimum observations and basin physical characteristics data and which cover a wide range of climate, soils and vegetation characteristics. (Where available, basins which have the additional data discussed in the following section will be selected to cover the range of characteristics.) The main types of required historical data are: hourly and daily gauged precipitation; daily maximum, minimum and average temperature; surface meteorological observations and daily average stream discharges. These minimum data requirements for MOPEX are actually quite modest, although a higher level of data would be desirable. Since the desirable level of data is unachievable for all basins, the most important requirement is the minimum level. The data requirements are summarized in Table 1.

The most basic minimum requirement is to have daily precipitation and streamflow with climatological monthly mean statistics of the following surface meteorological variables: air temperature; relative humidity; wind speed; and cloud cover. The surface observation statistics would be used to estimate potential evaporation for some schemes and radiative forcing for others. Experience in hydrological modelling is that good parameter estimates can be made with climatological statistics to estimate energy forcing.

Table 1 Summary of minimum basin required observations.

Description requirement	Minimum	Desired
Precipitation	Daily	Hourly
Streamflow	Daily	Hourly
Surface meteorology observations	Monthly statistics	Daily/Hourly

Basins from a wide range of climate regimes are required. Basins must be free of upstream flow regulation. Basins must have sufficient hydrometeorological observations (precipitation, temperature and streamflow). Some basins should have a strongly dominant soil type and a strongly dominant vegetation type. Data for a large number of basins are required.

Data for several hundred basins for a wide range of climate, soils, and vegetation regimes are believed to be needed to meet the MOPEX goal. Although it should be possible to do this globally, it will take many years to have a truly worldwide set of basins. To make the most of limited resources and to demonstrate results quickly, the MOPEX strategy for acquiring basin data is first to seek data from places where it is most readily available. By far the best single source of data is the USA, which covers a wide range of climate regimes, where there are high quality data, and where there are no restrictions on data distribution. Accordingly, more than 400 USA basins were identified that met the basin selection criteria. Then, data from additional basins, globally, can be used to test whether results from the USA basins are transferable to other basins throughout the world and to evaluate how much new information may be contained in data sets from other parts of the world. At the same time, steps have been and are being taken, to encourage countries and scientists throughout the world to contribute to the MOPEX database.

Required observations

Required observations include daily values of mean areal precipitation, mean areal maximum and minimum temperature, streamflow and climatological mean potential evaporation. Additional observations are desirable as discussed below.

A critical aspect of data set preparation to meet MOPEX objectives is to have research quality estimates of mean areal precipitation. A practical estimate of gauge density requirements was made by Schaake (1981) and Schaake *et al.* (2000) for river forecasting applications. The required number of gauges for a basin of area, A (km^2), is:

$$N = 0.6A^{0.3} \tag{1}$$

The exponent 0.3 implies that the required number of gauges doubles as the basin size increases by a factor of 10. The number of gauges given by this equation should give mean areal precipitation estimates for each time step that are accurate to within 20%, 80% of the time during thunderstorm rainfall events (in the 20 000 km^2 Muskingum River, Ohio basin). Equation (1) is reasonable to apply for basins between 200 and 20 000 km^2. The required number of gauges for basins of different size, according to equation (1), are given in Table 2.

Table 2 Desired minimum number of raingauges per basin.

Area (km^2)	Number of gauges
1	1
10	2
100	3
1000	6

Required basin characteristics

Supporting basin boundary, stream and land characteristics data relating to topography, soils and vegetation are also needed (Table 3). Some of these supporting data are available on an ISLSCP CD-ROM, but additional data and refinement of the ISLSCP data to a scale greater than 0.5° are required.

Table 3 Required basin physical characteristics.

Description requirement	Minimum	Desired
Elevation	5 km/5 m contours	1 km/1 m
Basin boundaries	10 km/location	1 km
Streams	10 km/location	1 km
Soils—texture, hydraulic properties, etc.	20 km	1 km
Vegetation—type, rooting depth, phenology, etc.	20 km/monthly	1 km/weekly
Geology	10 km	1 km

Desirable additional observations

Having actual measurements of meteorological surface variables at daily or hourly steps will improve the simulations of land surface schemes. If diurnal fluctuations of surface fluxes are to be simulated, detailed measurements of energy forcing variables are needed. These data are not critical to estimate those parameters that can be extracted from long periods of precipitation and runoff, although they might contribute to the development of improved parameter estimation techniques and to testing the techniques developed only with the minimum required data. Table 4 lists the desired additional observations.

Table 4 Summary of desired additional observations.

Description requirement	Minimum	Desired
Snow cover—satellite product	Seasonal statistics	Daily/1 km
Snow water equivalent	Seasonal statistics	Daily
Pan evaporation	Seasonal statistics	Daily
Clouds	Daily	3-hourly
Short wave radiation	Daily	Hourly
Long wave radiation	Daily	Hourly
Soil moisture	Weekly	Daily

DATA SET DEVELOPMENT FOR US BASINS

Streamgauges

The streamgauge data were selected from a subset of the US Geological Survey (USGS) streamgauge network. This subset includes most of the gauges in the USGS hydro-climatic data network (HCDN) (Slack *et al.*, 1992) or in a similar network selected by Wallis *et al.* (1991). Both of these networks include only gauges believed to be unaffected by upstream regulation and with long enough data records to be suitable for climate studies.

Basin boundaries

Basin boundaries were developed for each of the potential streamgauges from the subset of USGS gauges explained above. These boundaries were based on a DEM derived by the NOHRSC (National Operational Hydrologic Remote Sensing Center) (http://www.nohrsc.noaa.gov/). The NOHRSC provides and maintains the NWS Integrated Hydrologic Automated Basin Boundary System (IHABBS) GIS database to support river and flood forecasting throughout the nation. The IHABBS system uses a 15-arc second DEM that has been processed to have a hydrologically consistent connectivity of neighbouring grid points. The IHABBS connectivity files were used to create a hydrological connectivity file upstream from each of the potential stream gauges. Then basin boundaries were generated from the basin connectivity files. The location of each streamgauge in the DEM was adjusted slightly to get the best possible match between the digital boundary area and the total gauged area specified by the USGS. If the areas matched to within 10% the basin was accepted. Boundaries for several small basins could not be found. In the end a set of 1861 streamgauges were selected for potential use by MOPEX. The locations of these streamgauges are shown in Fig. 2.

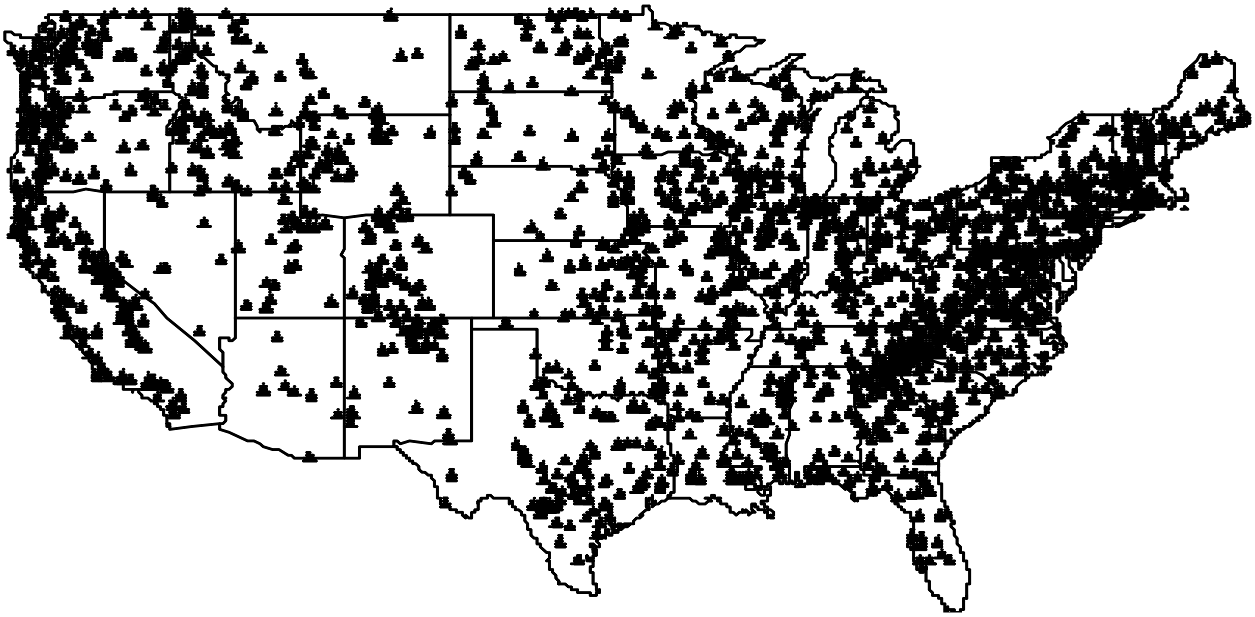

Fig. 2 Location of the 1861 potential MOPEX basins.

Precipitation observations

Hourly and daily precipitation data sets from 1948 to 2003 were assembled for the USA. Data sources included daily and hourly data sets from the National Climate Data Center (NCDC) (http://www.ncdc.noaa.gov/) and daily data from the Natural Resources Conservation Service (NRCS) SNOTEL network (http://www.wcc.nrcs.gov/factpub/-sntlfct1.html). Details about how these data were processed to create mean areal precipitation estimates are presented below.

Basin selection

Figure 3 compares the number of available gauges for each basin with the required number of gauges according to equation (1). Points that lie above the solid curve representing equation (1) have potentially sufficient data. Only 23% (438) of all basins have at least 80% of the required number of gauges. Figure 4 shows the locations of these 438 gauges. Subsequent analysis showed that adequate data for only 431 of the 438 stations were available.

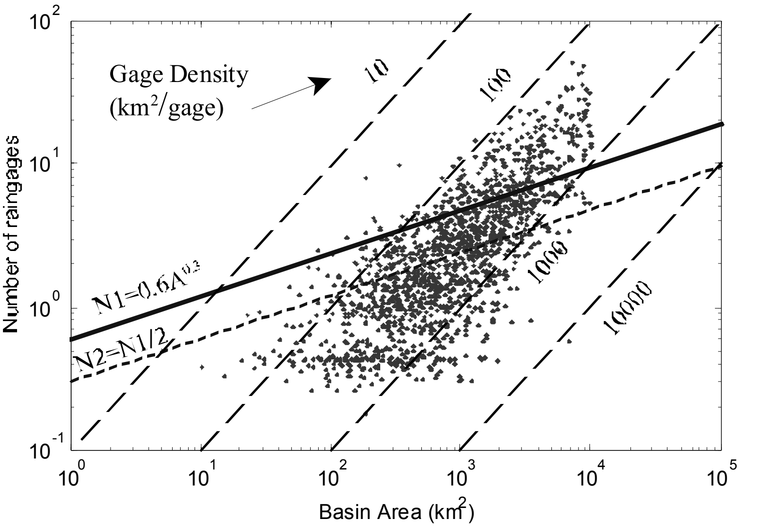

Fig. 3 Comparison of number of available gauges to number of required gauges at 1861 potential MOPEX basins.

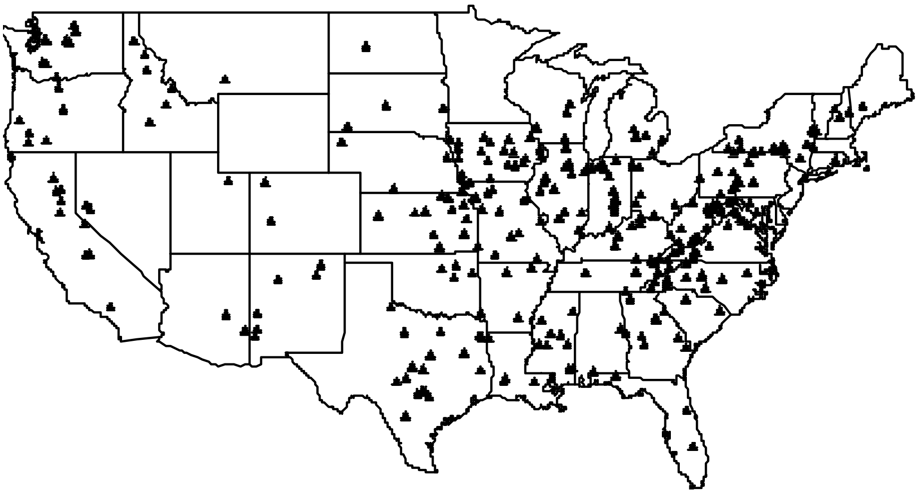

Fig. 4 Location of the 431 basins with an adequate number of precipitation gauges.

Basin characteristics

Gridded values of climate, soils and vegetation characteristics were used to derive basin characteristics for each basin. A useful variable to characterize the climate of each basin is the ratio of mean annual precipitation, P, to mean annual potential evaporation, EP. P/EP for each basin was estimated from gridded values of P from the PRISM project (http://www.ocs.orst.edu/prism_new.html) (Daly *et al.*, 1994) and gridded values of EP from the NOAA Evaporation Atlas (Farnsworth *et al.*, 1982). The distributions of P/EP values for basins within the Mississippi River basin are shown in Fig. 5. A number of approaches have been developed to classify vegetation and gridded files of these were processed to identify the vegetation distributions in each basin. These can be used to identify basins with very large fractions (say 80%) of each vegetation type. Soil hydraulic properties for each were developed from 1-km

STATSGO soil data provided by the Penn State Earth System Science Center (Miller *et al.*, 1999). STATSGO provides soil texture information. Hydraulic properties are derived from soil texture computed using several different empirical relationships (Clapp *et al.*, 1978; Cosby *et al.*, 1983). Figure 6 shows the distribution of basin average saturated hydraulic conductivity values derived for 39 basins in the Arkansas-Red River basin.

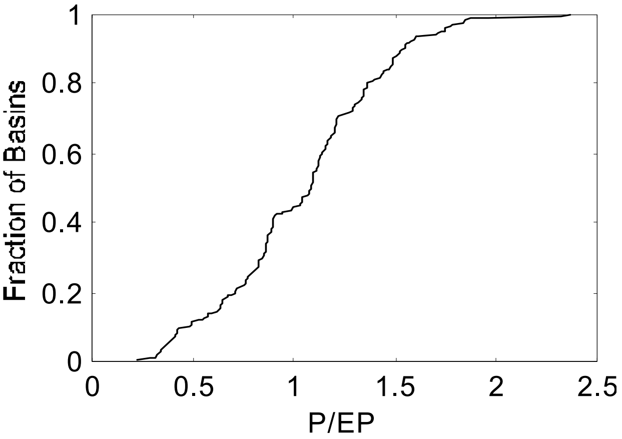

Fig. 5 Distribution of the ratio of mean annual precipitation to mean annual potential evaporation in the Arkansas-Red River basin.

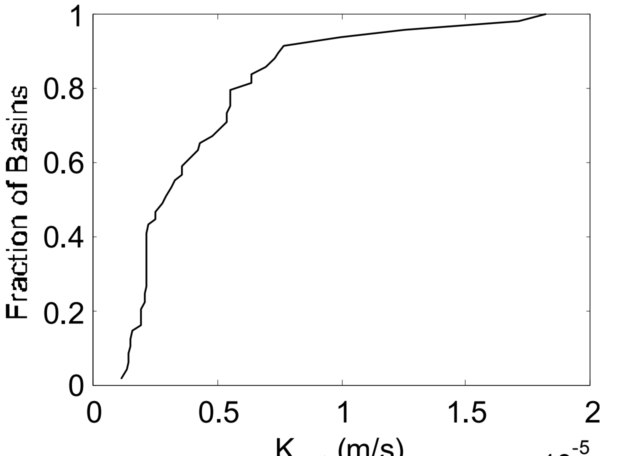

Fig. 6 Distribution of basin average saturated hydraulic conductivity in the Arkansas–Red River basin.

Soil texture The fractional spatial coverage of each of the 16 US Department of Agriculture (USDA) soil types (12 soil, 4 other category) was compiled for each basin. The analysis used 1-km gridded maps of USDA soil texture class for each of 11 soil layers. The 1-km gridded data sets were produced by Miller (1999) based on STATSCO polygon representation of soil texture. Table 5 provides the USDA soil classification definitions.

Table 5 The soil texture classification definitions.

1	S	Sand
2	LS	Loamy sand
3	SL	Sandy loam
4	SIL	Silt loam
5	SI	Silt
6	L	Loam
7	SCL	Sandy clay loam
8	SICL	Silty clay loam
9	CL	Clay loam
10	SC	Sandy clay
11	SIC	Silty clay
12	C	Clay
13	OM	Organic materials
14	W	Water
15	BR	Bedrock
16	O	Other

Table 6 The IGBP vegetation classification definitions.

1	Evergreen needleleaf forest
2	Evergreen broadleaf forest
3	Deciduous needleleaf forest
4	Deciduous broadleaf forest
5	Mixed forest
6	Closed shrublands
7	Open shrublands
8	Woody savannah
9	Savannahs
10	Grasslands
11	Permanent wetlands
12	Croplands
13	Urban and built-up
14	Cropland / natural vegetation mosaic
15	Snow and ice
16	Barren or sparsely vegetated
17	Water bodies

Table 7 The University of Maryland vegetation classification definitions.

0	Water (And Goode's Interrupted Space)
1	Evergreen needleleaf forest
2	Evergreen broadleaf forest
3	Deciduous needleleaf forest
4	Deciduous broadleaf forest
5	Mixed cover
6	Woodland
7	Wooded grassland
8	Closed shrubland
9	Open shrubland
10	Grassland
11	Cropland
12	Bare ground
13	Urban and built-up

Vegetation type Fractional coverage of each vegetation type according to each of two vegetation classification systems was compiled for each basin. The two vegetation classification systems were the International Geosphere-Biosphere Programme (IGBP) and University of Maryland (UMD). Table 6 gives the IGBP vegetation classes. Table 7 gives the UMD vegetation classes.

Greeness fraction Average monthly values of the fractional area coverage of vegetation were derived from NDVI data (Gutman & Iganatov, 1998). These data were originally compiled for most of North America Land Data Assimilation (NLDAS) project. Basin average values for each month were derived from the NLDAS grids for each of the 438 MOPEX basins.

Potential evaporation

Alternatives Experience in the calibration of hydrological models for river forecasting is that the year-to-year variability in the seasonal pattern of potential evaporation is not critical for model parameter estimation, although the effect of this can be seen in model performance and there is some influence on model parameters. One effect is that the active range of total water storage is about 10% greater if interannual variability of potential evaporation is accounted for and this affects the tuning of some model parameters. Some models do not use potential evaporation as an input and require the energy forcing data (e.g. SSIB and BATS). Energy forcing also is required for snow and frozen ground processes in some models. Energy forcing data sets for the US NLDAS domain on a 1/8 degree grid that were compiled by the University of Washington using empirical relationships between surface temperature and energy forcing variables (wind data were taken from the global reanalysis) were processed to produce basin-average energy forcing data sets. Comparisons between the 1/8 degree energy variable estimates and surface meteorological observations at 80 stations throughout the USA show that the temperature values are highly correlated but that other variables, especially wind, are not.

Climatology Climatological estimates of mean daily potential evaporation were made for each basin. The basis for these estimates is data from the NOAA Evaporation Atlas (Farnsworth *et al.*, 1982). This atlas contains maps of average annual and May–October free water surface potential evaporation. The NOAA Evaporation Atlas maps were digitised on a 1/6 degree grid. The NOAA Evaporation Atlas maps were derived by analysis of evaporation pan data. Pan evaporation was converted to free water surface evaporation using pan coefficients derived from studies of lake evaporation at several locations in the USA.

The annual cycle of mean potential evaporation was developed using average monthly pan data taken from the NOAA Evaporation Atlas. A Fourier series with only an annual cycle was fitted to evaporation pan monthly averages. This single frequency component accounted for almost all of the variance of the annual cycle of monthly average pan evaporation. Values of the maximum amplitude and phase angle were gridded. For each basin the amplitude and phase angle values were used together with values of the mean annual potential evaporation and mean May–October potential evaporation to estimate the basin average annual cycle of potential evaporation. Daily values of mean potential evaporation were produced by this analysis.

Temperature index Several methods exist to estimate potential evaporation from air temperature. One of these by Hargreaves (Jensen *et al.*, 1990), used air temperature as an index to energy budget terms in a combination equation (e.g. Penman). The Hargreaves equation uses daily maximum and minimum air temperature data. Therefore, daily mean areal maximum and minimum air temperature data were produced for each basin. The analysis technique is described below.

PRECIPITATION ANALYSIS

Priority was given to creating the highest possible quality basin-average hourly precipitation estimates. This required analysis of both hourly and daily precipitation data from NCDC since there are more than four times more daily precipitation gauges. The nearest hourly gauge was used to disaggregate the daily data to hourly estimates. This required daily gauge observation times to be used. About 1/3 of the observation times were missing, so a highly reliable correlation technique to estimate missing daily observation times was developed. Hourly mean areal precipitation estimates were made using the observed hourly and disaggregated hourly data. Also, a daily mean areal precipitation data set was created for a day ending at midnight so that the daily precipitation data are synchronized with the USGS daily streamflow data.

In the USA, MOPEX basins located west of longitude 100W are likely to be affected by orographic processes. This region includes 80 of the 438 total number of MOPEX basins. It also includes 35 of the 188 basins selected for streamflow verification as part of the NLDAS project. PRISM precipitation climatology data were used to assure that the estimated areal average precipitation estimates in the west preserved the PRISM climatology.

Observation times at NCDC daily COOP stations

Precipitation and maximum and minimum temperature data at NCDC daily COOP (co-operative network) stations are observed once per day at a specified time, e.g. 07:00 h. Different stations may have different observation times. Even for one station, the observation time may change from time to time during the period of record. The frequency at which different observations times occur for stations in Ohio are given in Table 8.

The total number of stations in Table 8 is 169. There are 90 stations where there is no observation time available at all. The major observation times are 7, 8, 18 and 24 hours.

It is essential to account for the effect of observation time in the computation of mean areal precipitation for the MOPEX basins. The strategy used to do this is to disaggregate the daily data to hourly using the time distribution of hourly precipitation at the nearest hourly gauge and using the observation time at the daily gauge.

It also is essential to consider on which calendar day the maximum and minimum temperatures occurred. Since minimum temperatures usually occur in the early morning, it is assumed that the minimum temperature occurred on the day of observation. But maximum temperatures typically occur in the afternoon. In that case it is likely that the maximum temperature for stations with AM observation times actually

Table 8 Frequency of Observation times of daily precipitation in Ohio.

Observation time (hours)	Duration in station-months	Frequency (%)
5	126	0.1
6	767	0.9
7	33635	37.7
8	30286	34.0
9	1073	1.2
10	109	0.1
11	435	0.5
13	20	0.0
16	195	0.2
17	3412	3.8
18	6539	7.3
19	2186	2.5
20	226	0.3
21	680	0.8
22	16	0.0
23	195	0.2
24	9256	10.4

occurred on the previous day. Therefore the station observation times were used to create a “local 24 hour” maximum daily temperature data set that was used to compute mean areal maximum temperature values.

Estimation of missing observation times at NCDC daily COOP stations

About one-third of the station observations times for NCDC daily COOP stations are missing from the digital NCDC station history records. These observation times may have a paper record at NCDC but there is no digital record of them. Because observation time is important to the MOPEX analyses, a method was developed to estimate missing observation times at precipitation stations and the same observation time was assumed to apply if the station also reported maximum and minimum temperature.

Table 9 Classification of the observation times of daily precipitation.

Class	Definition	Number of stations
I	percent of AM duration ≥ 80%	78
II	50%≤ percent of AM duration ≤ 79%	46
III	percent of PM duration ≥ 80%	14
IV	50%≤ percent of PM duration ≤ 79%	5
V	percent of 24 PM duration ≥ 80%	10
VI	50%≤ percent of 24 PM duration ≤ 79%	4
No observation time available at all		90
The rest		12
Total stations		259
Subtotal of stations in I to VI		157

Analysis of the daily COOP precipitation data for Ohio showed that the correlation coefficient for daily precipitation between two stations a given distance apart was strongest if the two stations had the same observation time. The correlation distances were found to be greatest during winter months.

The procedure to estimate missing observation times was first to classify station observation times as shown in Table 9.

Next, exponential distance decorrelation functions were developed for stations in each observation class and depending on the azimuth (eight compass points) of the line connecting the pair of stations. A set of these exponential decorrelation functions were developed for 5-degree latitude–longitude grid boxes for the contiguous USA.

It might be expected that the observation time for a given station could be inferred by comparing the correlation coefficients between the given station and its neighbouring stations with the expected correlation coefficient if the two stations had the same observations time class. The statistic used to make this comparison is the root-mean-square (RMS) difference for each observation time class between the correlation coefficients for the given station and the corresponding expected correlation coefficient for neighbouring stations in that class. The observation time for the given station was assumed to be given by the class with the minimum RMS difference.

Analysis of data for Ohio showed that the correlation coefficients in December and January were higher than that in other months. So January and December were selected for use. The analysis also showed that the averages of the correlation coefficients in azimuth sectors 5 and 6 were highest followed by averages in sectors 4 and 7. Therefore, sectors 5 and 6, and 4 and 7 were selected for use.

The reliability of the estimation method was tested for stations in Ohio. Observation times for stations with known observation times were estimated with the method using data from neighbouring stations. The results are given in Table 10. The rows in Table 10 correspond to the true observation time class. The columns corres–pond to the estimated observation class. The numbers in the table are the number of stations in a given class (row) that were classified in each of the other classes (column). If the technique was able to detect the correct class for every station, all of the numbers (>0) in Table 10 would fall on the diagonal.

The data in Table 10 suggest the technique is very reliable at detecting AM *vs* PM stations but not that reliable at determining the exact AM or PM time. Therefore, a set of three observation time classes was created by combining classes 1–2, 3–4 and 5–6. The results are in Table 11. Only a few values are off the diagonal so the technique seems to be highly reliable at detecting AM *vs* PM stations.

Table 10 Test of the observation class procedure using Sectors 5 and 6 in December for six classes.

	Estimation						
	59	18	1	0	0	0	78
T	27	13	2	4	0	0	46
r	0	1	10	2	0	1	14
u	0	0	2	2	0	1	5
e	0	0	2	0	5	3	10
	0	0	2	0	0	2	4
	86	32	19	8	5	7	157

Table 11 Test of the observation class procedure using Sectors 5 and 6 in December for three classes.

	Estimation			
T	117	7	0	124
r	1	16	2	19
u	0	4	10	14
e	118	27	12	157

Table 12 Test of the observation class procedure using Sectors 4 and 7, and 5 and 6 in January and December for three classes.

	Estimation			
T	120	4	0	124
r	0	17	2	19
u	0	2	12	14
e	120	23	14	157

By using inter-station correlations for more winter months and for more sectors, it might be possible to improve the reliability of the procedure. Accordingly the best results are given in Table 12 and the procedure used for Table 12 was adopted to estimate missing observation times.

It is not likely that stations with observations times at 24:00 h have missing observation times. Table 12 suggests that the procedure has a 95% reliability to detect AM *vs* PM observations times.

Construction of station data for different time periods

Several processed data sets were created to make the raw NCDC and SNOTEL data easier to use with hydrological models. These include:

(a) disaggregated daily to hourly NCDC COOP precipitation data;
(b) disaggregated daily to hourly SNOTEL precipitation data;
(c) daily 12z to 12z precipitation data;
(d) local 24 h daily precipitation data;
(e) local 6 h precipitation data;
(f) local 24 h maximum temperature for NCDC COOP stations.

Mean areal precipitation analysis

The approach used to estimate mean areal precipitation (MAP) for each of the MOPEX basins is essentially the same as used for gauge-only MAP estimation in the National Weather Service River Forecast System (NWSRFS) (www.nws.noaa.gov/oh/hrl/-nwsrfs/users_manual). The underlying interpolation procedure uses an inverse distance algorithm. The MAP procedure has the following steps:

(1) Get the basin boundary coordinates.

(2) Create an n by n analysis grid within the latitude/longitude window containing the basin. The value used for n is 30.
(3) Create a list of the N grid points within the basin boundaries.
(4) For each month of the year get the PRISM (Daly *et al.*, 1994) climatological mean precipitation value for each grid point in the basin (for the period 1961–1990).
(5) Create a station selection window with latitude/longitude limits a distance d_w outside of the latitude/longitude limits of the basin. The value used for d_w in this analysis is 50 km.
(6) For each time step that a MAP value is needed:
    (i) Get gauge precipitation data for each gauge in the station selection window.
    (ii) Get the PRISM climatological mean precipitation value for each gauge.
    (iii) Compute the ratio, f, of gauge precipitation to PRISM average precipitation for the current month.
(7) For each grid point:
    (i) Select the stations to be used to estimate precipitation at that grid point. Select the two nearest stations in each quadrant.
    (ii) Compute station weights to estimate precipitation at each grid point using inverse distance weighting with exponent m. The value used for m in this analysis is two.
    (iii) Apply the station weights for that grid point to gauge values of f to estimate f at the grid point.
    (iv) Multiply the grid point value of PRISM mean precipitation by the grid point estimate of f, to get the grid point estimate of precipitation.
    (v) Sum the grid point precipitation estimates and divide the sum by N.

This MAP procedure can be used for both daily and hourly time steps. This assures that daily totals of hourly MAP estimates are the same as estimated from an analysis of daily data if data for the same stations are used and the daily data at each station are equal to the sum of the hourly data.

The MAP program code was optimized to avoid computing station weights and getting PRISM values if the station list did not change or if the PRISM values had already been obtained. Also, the interpolation procedure was optimized by computing an effective station weight (implied by the procedure described above) that could be applied to each station to estimate MAP directly as a linear combination of gauge precipitation values. These weights were re-computed at the beginning of each month and whenever the station list changed within a given month.

The MAP procedure used for MOPEX is slightly different than the NWSRFS MAP procedure. The difference is that the value of f at the precipitation gauge is computed as the ratio of the gauge precipitation value to the PRISM mean value for the PRISM grid element where the gauge is located. The NWSRFS procedure uses the historical *station mean* (also called station normal in NWSRFS terminology) for the gauge, not the PRISM mean. This difference has the following implications:

(a) There must be enough data at a gauge to estimate the station normals for each month to use the NWSRFS procedure.
(b) The station normal may not be for the same climatological period as the PRISM mean value unless the period of record is the same or some method is used to account for the difference between the period gauged and the PRISM period.

(c) If the station normal is for the same period as the PRISM data, the NWSRFS estimation procedure would produce a mean precipitation estimate at a location very near the gauge that would equal the PRISM value, not the gauge mean value.

(d) If the station normal is not for the same period as the PRISM data, the NWSRFS estimation procedure would produce a mean precipitation estimate for the period of the gauge record at a location very near the gauge that would equal the PRISM value for gauge period, not the PRISM period.

The procedure used here to process the PRISM data has the following advantages:

(a) The MAP procedure used here can use precipitation data for any station without regard for its period of record.

(b) The MAP procedure used here will produce a mean precipitation estimate at a location very near a gauge that is equal to the gauge value.

(c) The mean value of estimated precipitation anomalies (relative to the PRISM grid) near gauges tend to be driven by the gauge anomaly (relative to the PRISM grid).

(d) The current procedure is less sensitive to the particular period of time used to estimate the PRISM climatology. It does not actually use the magnitude of the PRISM grid directly. It only uses the ratio of the PRISM value at one point to the PRISM value at another point. Although this ratio depends on the types of precipitation events that occurred over a period of time, it should not be very sensitive to differences in mean precipitation climatology over different periods of time. In other words, it might be expected that spatial ratios would be much less sensitive to the period chosen than the magnitude of the precipitation. As a result it is more important to update the PRISM grids if the PRISM analysis method improves than if a more recent period of data are used in the PRISM analysis.

(e) The existing NWSRFS procedure requires manual construction of precipitation isolines.

Because the quality of the gauge data used in the MAP analysis is very important and the quality of data for gauges that have operated for only short periods is problematic, only gauges with at least five years data are used in this analysis.

Another difference between the MAP procedure used here and the MAP procedure used by the NWSRFS is that the current procedure does not estimate missing data for a station. The NWSRFS procedure first estimates missing station data so there is a data value for every station for the entire period of the analysis. The NWSRFS procedure can be shown not to modify the effective weights applied to the non-missing data to get the MAP value. The NWSRFS procedure is easier to program than the current procedure and is only slightly more computationally efficient.

TEMPERATURE ANALYSIS

Temperature data sources

Data sources for temperature analysis were daily maximum and minimum temperatures at NCDC COOP stations and at SNOTEL sites.

Estimation of missing temperature observation times

The observation times estimated for missing temperature observation times were assumed to be the same as for the corresponding precipitation observation.

Mean areal temperature analysis

The procedure to estimate mean areal temperature involved first estimating mean areal maximum and minimum daily temperature. Then, if temperature values during the day are needed, the procedure suggested by Parton & Logan (1981) to estimate hourly temperatures from the daily maximum and minimum was used.

US DATA SETS

Several different kinds of data sets were produced for the USA by the MOPEX project. These are available by anonymous ftp at ftp://hydrology.nws.noaa.gov/pub/gcip/-mopex/US_Data/. Gauge data were obtained for the period of record for each station for the period 1948 to the present. A general description of the data available at this site follows. More detailed documentation is available on the ftp site.

Mean areal precipitation and temperature data are organized as time series data for each basin. A composite daily time series data set was also produced for each basin that contains daily mean areal values for precipitation, potential evaporation, streamflow, maximum temperature and minimum temperature.

The gauge-based data sets used to generate the mean areal time series are organized in a "random-spatial" format. Daily gauge-based data are organized so there is a data file for each month. This file contains a record for each station with data for at least one day of that month. All stations for the contiguous USA are in that file. The monthly files are organized into directories for the decades 1940, 1950, …, 2000. Hourly gauge-based data are organized so there is a file for each day with a record for each station containing 24 hourly values for that day. The daily data files are organized into monthly subdirectories.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/US_438_Daily
This directory contains a time series daily data file of each of the MOPEX basins that met the minimum precipitation gauge density requirements. There are data for 431 basins in this directory.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/Basin_Characteristics
This directory contains files and subdirectories with basin characteristics data for each of the 431 basins with data subdirectory US_438_Daily.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/Basin_Boundaries
This directory contains basin boundary data files for each of the 431 basins with data subdirectory US_438_Daily.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/Hourly_MAP
This directory contains hourly MAP files for each of the 431 basins with data subdirectory US_438_Daily.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/6hr_MAP
This directory contains 6-hourly MAP files for each of the 431 basins with data subdirectory US_438_Daily.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/6hr_MAT
This directory contains 6-hourly MAT files for each of the 431 basins with data subdirectory US_438_Daily.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/Daily_Q_1800
This directory contains time series files of daily streamflow for each of the 1862 gauges that were potential US MOPEX basins.

hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/Ameriflux Data
This subdirectory contains information about the Ameriflux network, including data sets for some stations.

hydrology.nws.noaa.gov/pub/gcip/station_data/precipitation/dpndata/
This directory contains subdirectories for the following gauge-based daily precipitation data sets:

(a) ncdc_24lcl—NCDC daily station data processed using hourly data to be on a local 24-hour clock. This subdirectory contains separate subdirectories for data from daily and hourly stations. Each of these subdirectories contains a subdirectory for each decade from 1940 to 2000.
(b) ncdc_ 12z—NCDC daily station data processed using hourly data to be on a daily 12z to 12z clock. The date corresponds to the day at the end of the 12z valid observation period. This subdirectory contains separate subdirectories for data from daily and hourly stations. Each of these subdirectories contains a subdirectory for each decade from 1940 to 2000.
(c) snotel_24lcl—original daily SNOTEL data on a local 24-hour clock. This subdirectory contains subdirectories for each decade 1970 to 2000.
(d) snotel_12z—SNOTEL daily station data processed using hourly data to be on a daily 12z to 12z clock. The date corresponds to the day at the end of the 12z valid observation period. This subdirectory contains subdirectories for each decade 1970 to 2000.

hydrology.nws.noaa.gov/pub/gcip/station_data/precipitation/hpndata/
This directory contains subdirectories for the following gauge-based hourly precipitation data sets:

(a) ncdcz—NCDC hourly data on a UTC 0–24z clock. These data are from hourly gauges.
(b) ncdc_ disagg—disaggregated NCDC daily data to hourly on a UTC 0–24z clock.

(c) snotel_disagg—disaggregated hourly snotel daily data on a UTC 0–24z clock.

hydrology.nws.noaa.gov/pub/gcip/station_data/temperature/
This directory contains subdirectories for the following gauge-based daily station data sets:

(a) snoteldtmn—SNOTEL daily minimum temperature.
(b) snoteldtmx—SNOTEL daily maximum temperature.
(c) tmax_24lcl—NCDC daily maximum temperature (observation time adjusted).
(d) tmin—NCDC daily minimum temperature.

Acknowledgements The NOAA Office of Global Programs provided financial support under the GEWEX Americas Prediction Project (GAPP) program for development of the US basin data set for the MOPEX database.

REFERENCES

Clapp, R. B. & Hornberger, G. M. (1978) Empirical equations for some soil hydraulic properties. *Water Resour. Res.* **14**(4), 601–604.

Cosby, B. J., Hornberger, G. M. Clapp, R. B. & Ginn, T. R. (1984) A statistical relationship of soil moisture characteristics to the physical properties of soils. *Water Resour. Res*. **20**, 682–690.

Daly, C., Neilson, R. P. & Phillips, D. L. (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Met.* **33**, 140–158.

Farnsworth, R. K., Thompson, E. S. & Peck, E. L. (1982) Evaporation Atlas for the contiguous 48 United States. *NOAA Technical Report, NWS 33, Washington, DC, USA.*

Franz, K., Ajami, N., Schaake, J. & Buizza, R. (2005) Hydrologic Ensemble prediction experiment focuses on reliable forecasts. *Eos, Trans. AGU* **86**(25), 239.

Gutman, A. & Iganatov, A. (1998) The derivation of the green vegetation fraction from NOAA/AVHRR data for use in numerical weather prediction models. *Int. J. Remote Sens.* **19**(8), 1533–1543.

Hogue, T., Wagener, T., Schaake, J., Duan, Q., Hall, A., Gupta, H. V., Leavesley, G. & Andreassian, V. (2004) A new phase of the Model Parameter Estimation Experiment (MOPEX). *Eos, Trans. AGU* **85**(22), 217–218.

Jensen, M. E., Burman, R. D. & Allen, R. G. (1990) Evapotranspiration and irrigation water requirements. *ASCE Manual 70*, 232.

Miller, D. A. & White, R. A. (1999) A conterminous United States multi-layer soil characteristics data set for regional climate and hydrology modeling. *Earth Interactions 2* (available at http://EarthInteractions.org).

Parton, W. J. & Logan, J. A. (1981) A model for diurnal variation in soil and air temperature. *Agric. Met.* **23**, 205–216.

Schaake, J. C. (1981) Summary of river forecasting raingauge network density requirements (unpublished). Available at http://www.nws.noaa.gov/oh/mopex/raingauge%20density%20requirement.htm.

Schaake, J. C., Duan, Q., Smith, M. & Koren, V. (2000) Criteria to select basins for hydrologic model development and testing. Preprints in: *15th Conf. On Hydrology* (Long Beach, California, USA, Am. Met. Soc., 10–14 January 2000), Paper P1.8.

Slack, J. R. & Landwehr, J. M. (1992) Hydro-climatic data network (HCDN): A US Geological Survey streamflow data set for the United States for the study of climate variations, 1874–1988. *USGS Open-file Report 92-129, Reston, Viginia, USA.*

Wallis, J. R., Lettenmaier, D. P. & Wood, E. F. (1991) A daily hydroclimatological data set for the continental United States. *Water Resour. Res*. **27**(7): 1657–1663.