INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **25**: 649–664 (2005)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/joc.1144

SEASONAL FORECASTING OF THAILAND SUMMER MONSOON RAINFALL

NKRINTRA SINGHRATTNA,[a,b] BALAJI RAJAGOPALAN,[a,c,*] MARTYN CLARK[c] and K. KRISHNA KUMAR[c,d]

[a] *Department of Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, CO, USA*
[b] *Thailand Public Works Department, Bangkok, Thailand*
[c] *Co-operative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA*
[d] *Indian Institute of Tropical Meteorology, Pune, India*

Received 19 April 2004
Revised 3 November 2004
Accepted 3 November 2004

ABSTRACT

This paper describes the development of a statistical forecasting method for summer monsoon rainfall over Thailand. Predictors of Thailand summer (August–October) monsoon rainfall are identified from the large-scale ocean–atmospheric circulation variables (i.e. sea-surface temperature and sea-level pressure) in the Indo-Pacific region. The predictors identified are part of the broader El Niño southern oscillation (ENSO) phenomenon. The predictors exhibit a significant relationship with the summer rainfall only during the post-1980 period, when the Thailand summer rainfall also shows a relationship with ENSO. Two methods for generating ensemble forecasts are adapted. The first is the traditional linear regression, and the second is a local polynomial-based nonparametric method. The associated predictive standard errors are used for generating ensembles. Both the methods exhibit significant comparable skills in a cross-validated mode. However, the nonparametric method shows improved skill during extreme years (i.e. wet and dry years). Furthermore, the models provide useful skill at 1–3 month lead time that can have a strong impact on resources planning and management.

KEY WORDS: Thailand; summer rainfall; monsoon; ENSO; ensemble forecast; nonparametric methods; local polynomials; seasonal forecasting

1. INTRODUCTION

Seasonal forecasts of Thailand summer monsoon rainfall can have significant value for resources planning and management, e.g. reservoir operations, agricultural practices, and flood emergency responses. In particular, increased population stress on the Chao Phraya River basin, one of the key regions for Thailand's socio-economic well being, is resulting in water quantity and quality problems. To mitigate this, effective planning and management of water resources is necessary. In the short term, this requires a good idea of the upcoming monsoon season rainfall, i.e. a good seasonal forecast. In the long term, it needs realistic projections of scenarios of future variability and change. There is no known long-lead forecast of Thailand summer monsoon rainfall or stream flows. As a result, much of the water resource planning in Chao Phraya basin and in Thailand in general is near term, i.e. responding to near-term weather forecasts.

There is an extensive literature of studying the variability of Indian summer monsoon, both from observational data (e.g. Walker, 1924; Pant and Parthasarathy, 1981; Rasmusson and Carpenter, 1983; Fein and Stephens, 1987; Webster *et al.*, 1998) and from modelling studies (e.g. Ju and Slingo, 1995; Meehl and Arblaster, 1998). These studies have identified a strong link between El Niño southern oscillation (ENSO) and the Indian summer monsoon. Statistical methods for forecasting the Indian monsoon rainfall

* Correspondence to: Balaji Rajagopalan, Department of Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, Co 80309-0428, USA; e-mail: balajir@spot.colorado.edu

use this ENSO–monsoon relationship. For example, Krishna Kumar *et al.* (1995) and Shukla and Mooley (1987) identified several predictors of the Indian monsoon and developed statistical models for forecasting — almost all of the predictors are various facets of ENSO. With this framework of predictors, statistical models using time series (Thapliyal, 1981; Rajeevan, 2001) and artificial neural network techniques (Sahai *et al.*, 2000) have been developed by the Indian Meteorological Department and other researchers for use in operational forecasts.

Krishna Kumar *et al.* (1999a) showed that the ENSO–Indian monsoon relationship has weakened substantially in the post-1980 period. They argue for changed ENSO characteristics and global warming as potential causes for this weakening. This is having a strong impact on the forecasting efforts of the Indian monsoon, as most of its predictors (mentioned above) are related to ENSO. Furthermore, Krishna Kumar *et al.* (1995, 1999b) show that the Indian monsoon predictors are strongly related to the Indian monsoon only when the monsoon itself is strongly related with ENSO. Interestingly, results from our research (Singhrattna *et al.*, in press) indicated that the Thailand monsoon is more closely related to ENSO in the post-1980 period, just when the Indian monsoon relationship with ENSO is weakening. This enhances the prospects of forecasting Thailand monsoon rainfall.

There is little in the literature studying the variability and predictability of the Thailand summer monsoon. Admittedly, it is much smaller in comparison with the Indian summer monsoon, but it has a significant socio-economic impact in Thailand. There have been some studies of late on the variability of the Thailand monsoon and rainfall over Singapore and Indonesia by Kripalani and Kulkarni (1997, 1998, 2001) and more recently by Singhrattna (2003; Singhrattna *et al.*, in press). Distributed hydrologic models for Chao Phraya and Nakon Sawan river basins have been developed (Jha *et al.*, 1997, 1998), but these are mainly for real-time or event-based simulation of stream flow and not for seasonal forecasting.

It is not clear whether there is a seasonal forecast mechanism in place in Thailand. Unlike in the case of the Indian summer monsoon, the Indian Meteorological Department is required to issue a seasonal forecast of the upcoming monsoon season by the end of April. The great need and utility of a Thailand monsoon forecast and the enhanced prospects of its predictability in recent decades serve as a strong motivation for the present research.

We adapt two approaches for an ensemble forecast of Thailand summer monsoon rainfall in this paper. The first is a traditional linear regression approach and the second a nonparametric technique based on local regressions.

The paper is organized as follows. The data description and predictor identification are presented first. The two forecasting methods are then described, followed by cross-validated model skills in forecasting the Thailand summer rainfall. A discussion of the results concludes the paper.

2. DATA

The data used in this study are:

1. Rainfall data for the Thailand summer monsoon (August–October), and surface air temperature (SAT) data during pre-monsoon months (March–June) averaged over three stations: Nakon Sawan (15°48′N, 100°10′E), Suphan Buri (14°28′N, 100°08′E) and Don Muang (13°55′N, 100°36′E). All of these stations are in the west central region and in the Chao Phraya River basin. These data were obtained from the GEWEX Asian Monsoon Experiment (GAME) project Website (http://hydro.iis.u-tokyo.ac.jp/GAME-T/GAIN-T/routine/rid-river/longterm.html). The GAME programme, part of the global energy and water cycle experiment (GEWEX), has done a good job of collecting and archiving data from South East Asian countries. In general, it has been difficult obtaining long hydroclimate data from South East Asia, and Thailand in particular. See Singhrattna (2003) for further details on these data sets.
2. Large-scale ocean and atmospheric circulation variables, such as sea-surface temperature (SST), sea-level pressure (SLP), winds, velocity potential, were obtained from National Center for Environmental Prediction–National Center for Atmospheric Research reanalysis (Kalnay *et al.*, 1996). These data sets

span the period from 1948 to date, covering the globe on a $2.5° \times 2.5°$ grid and are available at http://www.cdc.noaa.gov.

3. Standard ENSO indices: NINO3, NINO1 + 2, southern oscillation index (SOI) available at http://www.cpc.noaa.gov.
4. Indian Ocean dipole (IOD) index (Saji *et al.*, 1999). This is an index based on the SST anomaly difference between the eastern and western tropical Indian Ocean. The index, its impact on the adjoining continental rainfall, interactions with ENSO and teleconnections can all be obtained from the IOD home page http://www.jamstec.go.jp/frsgc/research/d1/iod/.

3. IDENTIFICATION OF PREDICTORS

The aim in this section is to identify predictors for the Thailand summer rainfall, which can then be used in statistical forecast models. The two main requirements for any useful predictors are (1) a good relationship with the seasonal rainfall and (2) a reasonable lead time (i.e. months to a season). Our earlier work (Singhrattna, 2003; Singhrattna *et al.*, in press) indicated that Thailand summer rainfall is strongly correlated with ENSO in the post-1980 period and also with pre-monsoon (especially March–May (MAM)) land surface temperatures representing the land–ocean thermal gradient. So, the first step is to look for a relationship with standard ENSO indices during the pre-monsoon seasons and follow up with correlations between the rainfall and large-scale ocean–atmospheric variables (SSTs, SLPs). This approach of correlation with large-scale ocean–atmospheric circulation variables has been used to identify predictors for stream flows in northern Brazil (Souza and Lall, 2003) and in the Truckee–Carson river basins in Nevada, USA (Grantz, 2003).

3.1. Correlation with ENSO indices

Thailand summer monsoon rainfall was correlated with the standard ENSO indices and IOD index from pre-monsoon seasons and also with the spring (MAM) Thailand air temperatures (SAT). The latter are believed to be an indicator of the land–ocean thermal gradient, which is important for the strength of the monsoon (Singhrattna *et al.*, in press). The correlations are computed for the post-1980 period and shown in Table I. Correlation values that are statistically significant at the 95% confidence level using a t-test (Helsel and Hirsch, 1995) are shown in bold in the table. It can be seen that SOI, the SLP-based ENSO index, shows a strong correlation with monsoon rainfall during the concurrent season and also one or two seasons prior. The spring land temperatures also exhibit a significant correlation, as expected. The IOD shows a strong correlation with the monsoon rainfall at a two-season lead time. All these brighten the prospects for a long-lead forecast.

To confirm that the correlations are strong only during the post-1980 period (as in Table I), selected predictors from pre-monsoon seasons (January–March (JFM) NINO3, May–July (MJJ) SOI, MAM IOD and MAM SAT) were correlated with monsoon rainfall on a 21 year moving window (Figure 1). It can be seen that the predictors show correlations with summer rainfall only in recent decades, much as the correlations between the rainfall and ENSO (shown as a solid line between summer rainfall and ASO SOI) and seen by Singhrattna *et al.* in press. Similar shifts have been seen (Miyakoda *et al.*, 2003) in pre-monsoon signals of

Table I. Correlations (post-1980 period) between Thailand summer rainfall (August–October) and large-scale climate indices (the 95% significance level is ±0.41. Values in bold are statistically significant at the 95% level)

	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO
Nino 1 + 2	**0.41**	0.31	0.29	0.28	0.25	0.17	0.08	−0.06
Nino 3	**0.42**	0.33	0.15	−0.01	−0.13	−0.19	−0.24	−0.31
SOI	0.40	0.27	−0.07	−0.27	**0.44**	**0.45**	**0.57**	**0.59**
IOD	−0.37	**−0.44**	**−0.70**	**−0.55**	−0.32	−0.17	−0.22	−0.34
SAT	0.30	**0.51**	**0.48**	0.34	0.20	0.10	−0.01	−0.11

the South Asian monsoon. This suggests that the ENSO-based predictors are related to the monsoon rainfall only when the monsoon rainfall itself is related to ENSO. Interestingly, this is similar to the finding by Krishna Kumar *et al.* (1995), who show that the predictors of Indian monsoon rainfall are related to the rainfall only during the period when the Indian monsoon is strongly related with ENSO. In the case of the Indian monsoon this is the pre-1980 period. This is consistent with the ENSO-related circulation changes during pre- and post-1980 periods (Krishna Kumar *et al.*, 1999a; Singhrattna *et al.*, in press). Land-cover changes (Kanae *et al.*, 2001) and decadal changes in the ENSO–monsoon relationship (Krishna Kumar *et al.*, 1999b; Torrence and Webster, 1999) could lead to trends in monsoon precipitation and, consequently, add to the nonstationarity of the relationship, as seen in Figure 1.

3.2. *Correlation with large-scale variables*

Although, as seen above, the indices show significant correlations, we would like to check their large-scale aspects and also see whether other, stronger predictors could be identified. To this end, the summer monsoon rainfall was correlated with SSTs and SLPs during pre-monsoon seasons and the correlation maps are shown in Figure 2. The shaded regions indicate correlations that are significant at the 95% confidence level. With SLPs (Figure 2(a)) the correlations are strong in the Pacific subtropical region, indicating that a higher than normal subtropical pressure tends to enhance the easterlies, thereby increasing the moisture transport to Thailand and, consequently, the rainfall. Wang *et al.* (2003: figures 1 and 2) found similar pressure patterns in the Pacific subtropical region to be linked with variations in the Australian and Asian monsoons. Strong positive correlations with SSTs (Figure 2(b)) are seen in the eastern Indian Ocean and western Pacific Ocean regions around the equator. This region is also one of the poles of the IOD index (Saji *et al.*, 1999); hence the strong correlation with IOD seen in Table I and Figure 1. These correlation maps indicate persistence from the spring leading up to the monsoon season, thus providing the potential for a long-lead forecast. The solid boxes in Figure 2 show the regions of high correlation from where the predictors will be developed in the following sections.

3.3. *Predictor selection*

Based on the correlations with indices and the correlation maps with large-scale variables, predictors with high correlations with the summer rainfall were identified. With this criterion, the selected predictors are (1) SSTs averaged over 10.5–14.5 °S latitudes and 108–120 °E longitudes and (2) SLPs averaged over

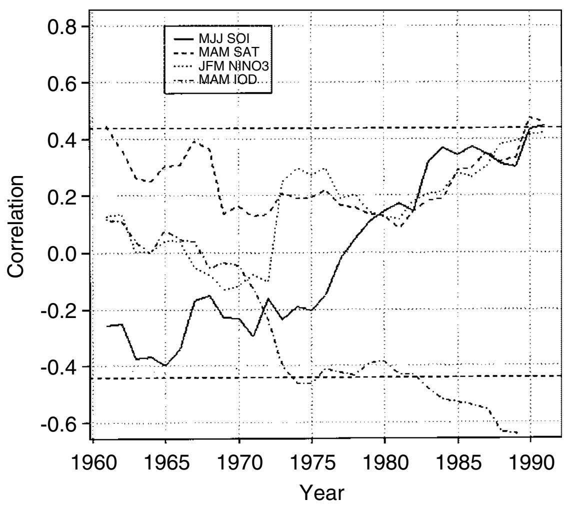

Figure 1. 21-year moving window correlation between Thailand summer (August–October) rainfall and selected predictors from pre-monsoon seasons (JFM NINO3; MAM IOD; MAM SAT; May–July SOI). The dashed horizontal lines are the 95% significance levels

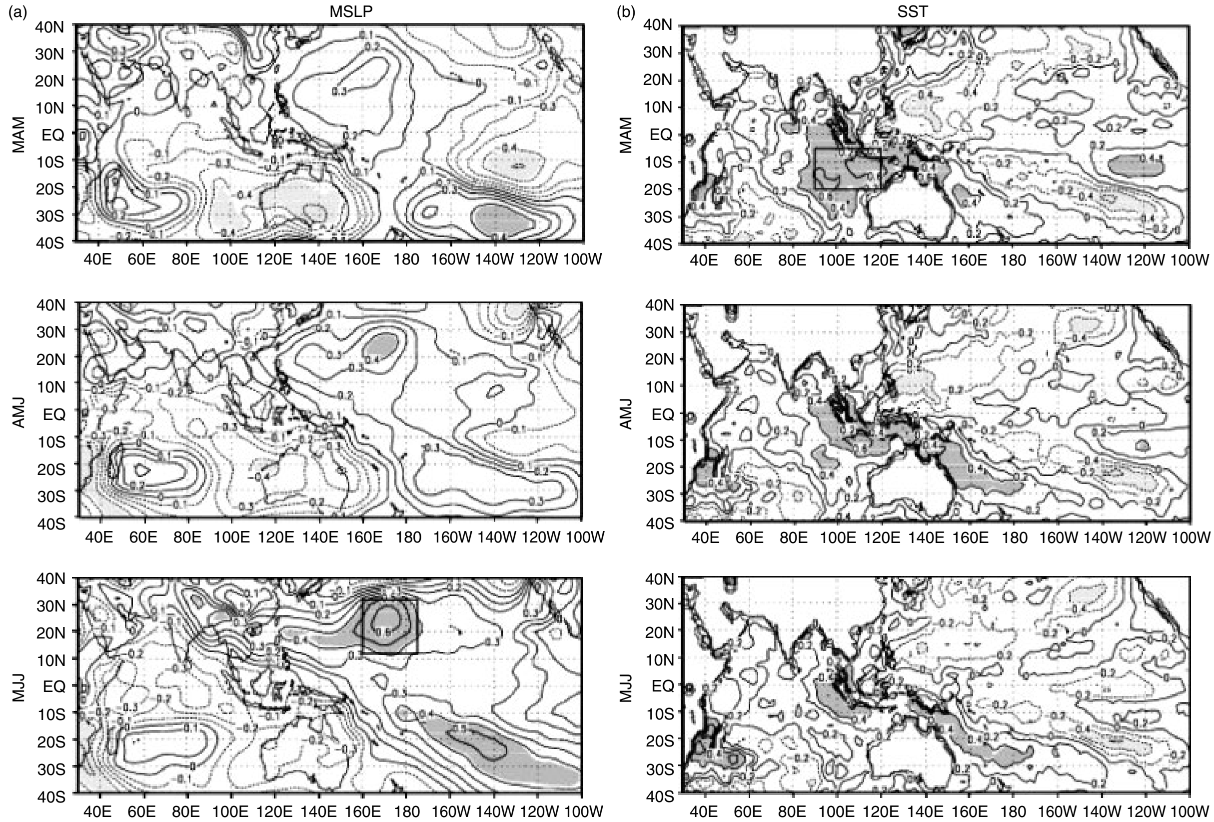

Figure 2. Correlation maps of Thailand summer rainfall and pre-monsoon season: (a) SLPs; (b) SSTs. Shaded regions are significant at the 95% confidence level

20–30 °N latitudes and 165–180 °E longitudes. Thailand SAT is also selected as one of the predictors. This essentially captures the land–ocean gradient that gets set up by the land temperatures, especially during the spring season before the monsoon (Singhrattna, 2003).

To check the temporal variability of the strength of the predictors to monsoon rainfall, moving window correlations are shown in Figure 3. As expected, the predictors are correlated mainly in the post-1980 period as in Figure 1. Furthermore, the predictors show significant correlations with the summer rainfall at a one- to two-season lead time.

4. FORECAST MODELS

Typically, a regression (often linear) is fit between the identified predictors and a single dependent variable (i.e. the summer rainfall). The fitted regression is then used to forecast the mean value of the variable. There is an extensive literature for fitting and testing linear regression models, and software is readily available (e.g. Helsel and Hirsch, 1995). Such models have been widely used for hydroclimate forecasting in the USA (e.g. Lui *et al*., 1998; Cordery and McCall, 2000; Piechota *et al*., 2001; McCabe and Dettinger, 2002) and for Indian monsoon forecasting (Hastenrath, 1987, 1988; Krishna Kumar *et al*., 1995). For forecasting a field of a dependent variable, such as precipitation, at several locations from fields of independent variables (e.g. tropical SST, SLP, etc.), canonical correlation analysis is typically used (e.g. Shabbar and Barnston, 1996; Ntale *et al*., 2003). Below, the linear regression model is briefly described.

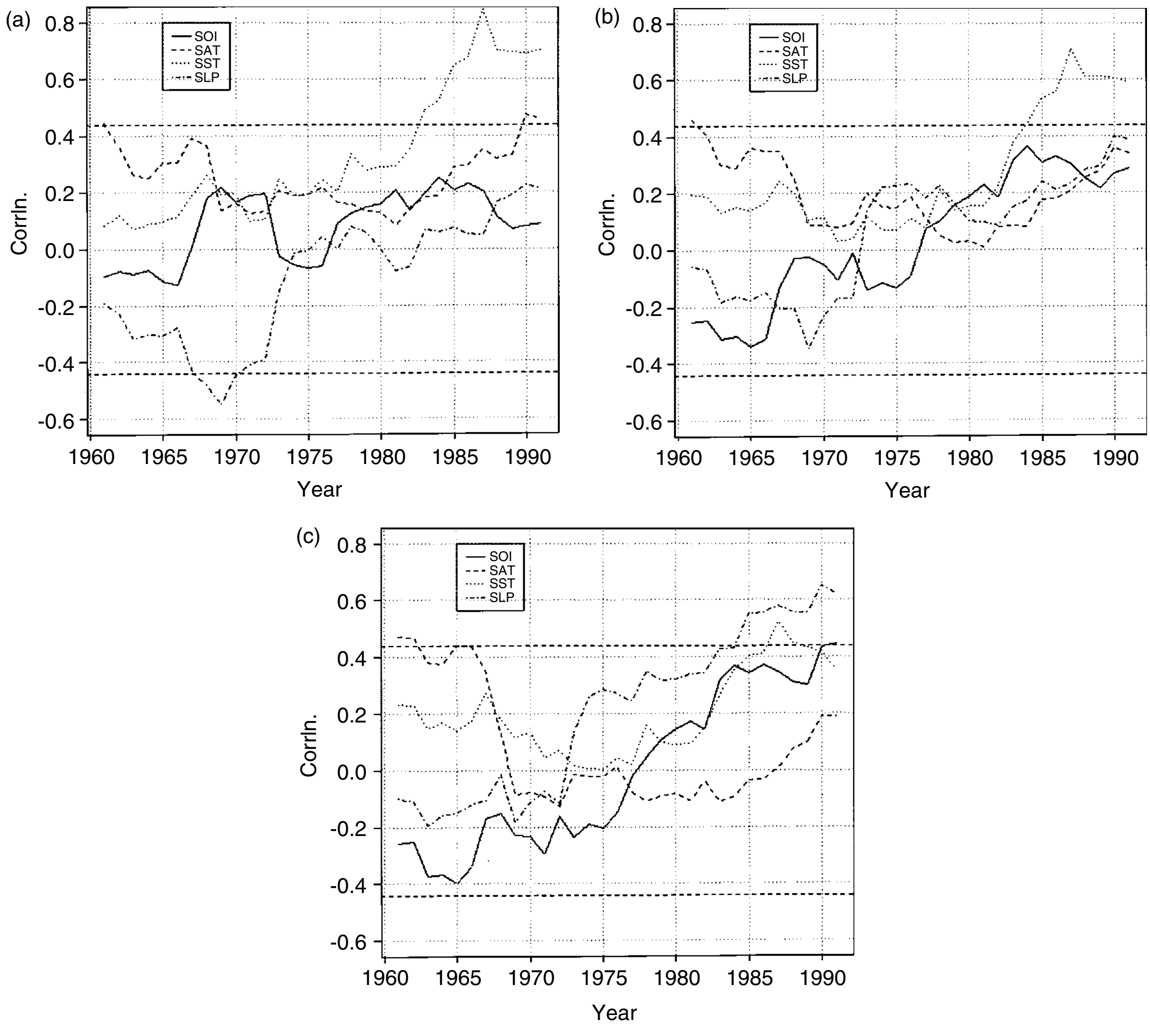

Figure 3. Same as Figure 1, but with the predictors identified for the pre-monsoon seasons: (a) MAM; (b) AMJ (c) MJJ

4.1. Linear regression

Traditional linear regression involves fitting a linear function between the response variable (i.e. summer rainfall) and the independent variables (i.e. predictors). They are of the form

$$Y_t = a_1x_{1t} + a_2x_{2t} + a_3x_{3t} + \ldots + \mathrm{a}_p{}^*x_{pt} + e_t \qquad t = 1, 2, \ldots, N \tag{1}$$

where the coefficients a_1, a_2, ..., a_p are estimated from the data, by typically, minimizing the sum of squares of the errors; e_t is the error, which is assumed to be normally (or Gaussian) distributed with zero mean and variance σ_e^2 (also estimated from the data); N is the number of observations. The equations for the coefficients, the error variance and methods for testing the goodness of the fitted model can be found in any standard book on statistics (e.g. Helsel and Hirsch, 1995).

Implicitly, the variables are also assumed to be normally distributed. If not, they are generally transformed to a normal distribution (e.g. log or power transform) before the model is fit. Once the model is fit (i.e. the coefficients estimated) then, for any new value of the predictors, the model with the fitted coefficients (Equation (1)) is used to predict the mean value of the dependent variable, say Y_{new}. Predictive standard error σ_{pe} (or the standard deviation of the error of the predicted mean) is obtained from theory (Helsel and Hirsch, 1995). Normal random deviates with a zero mean and standard deviation σ_{pe} provide the ensembles of errors, which when added to the mean estimate Y_{new} results in the ensemble forecasts. This approach of using a normal distribution with the predictive standard error was applied by Clark and Hay (2004) for generating ensemble forecasts of stream flows in the western USA.

In the above model, if the independent variables happen to be past values of the response variable itself, then it forms a time series model in an autoregressive framework. Hydrologists have developed and used such models for stream flow simulation and forecasting (Bras and Iturbe, 1985; Salas, 1985; Yevjevich, 1972).

The main drawbacks of traditional linear regression models are (1) the assumption of a Gaussian distribution of data and errors, (2) the assumption of a linear relationship between the variables, and (3) the models are not portable across data sets (i.e. sites). Furthermore, if the model fitted is found to be inadequate then the alternative choices are limited, and more so when the number of observations are small.

4.2. Nonparametric regression: locally weighted polynomials

Nonparametric methods provide an attractive alternative in alleviating some of the drawbacks of the traditional linear regression. In this approach, the model is

$$Y_t = f(x_t) + e_t \tag{2}$$

where $x_t = (x_{1t}, x_{2t}, x_{3t}, \ldots, x_{pt})$, $t = 1, 2, \ldots, N$. This is similar to the linear regression model (Equation (1)), but the function f could be linear or nonlinear, and the errors e_t are assumed to be normally distributed with zero mean and variance σ^2_{le}. The key difference from linear regression is that the function f is fit 'locally' to estimate Y. In that, the value of the function at any point x_i is obtained by (1) identifying a small number K $(= \alpha N$, where $\alpha \in (0, 1])$ of neighbours to x_i and (2) fitting a polynomial of order p to the neighbours. Neighbours are identified from the observations that are closest to x_i in terms of the Euclidian distance or another such metric (e.g. Mahalanobis distance; Yates *et al.*, 2003). The fitted polynomial is then used to estimate the mean value of the dependent variable. The coefficients of the polynomial are estimated using a weighted least-squares approach. The theoretical background of the local polynomial method is described in detail in Loader (1999), who refers to it as LOCFIT; henceforth, we will use the same terminology in this paper.

LOCFIT also provides the local standard errors of the estimate σ_{le} and local predictive standard errors σ_{lpe} (Loader, 1999), corresponding to σ_{e} and σ_{pe} respectively in the case of linear regression described in Section 4.1. The steps for generating the ensembles are the same as that for the linear regression: (1) for a new value of the predictor set, the mean value Y_{new} is first estimated using the LOCFIT approach described above; (2) the local predictive standard error σ_{lpe} is estimated (Loader, 1999); and (3) normal random deviates with a zero mean and standard deviation of σ_{lpe} when added to the mean estimate Y_{new} result in ensemble forecasts.

The key parameters to be estimated are the size of the neighbourhood (K or α) and the order of the polynomial p. These parameters are obtained using objective criteria such as the generalized cross-validation (GCV) function or likelihood function:

$$\mathrm{GCV}(\alpha, p) = \frac{\displaystyle\sum_{i=1}^{N} \frac{e_i^2}{N}}{\left(1 - \dfrac{m}{N}\right)^2} \tag{3}$$

where e_i is the error (i.e. difference between the model estimate and observed), N is the number of data points and m is the number of parameters. For a suite of α and p values the GCV function is computed from Equation (3) and the combination that gives the least GCV value is selected. For stability purposes, the minimum neighbourhood size should be twice the number of parameters to be estimated in the model.

Note that if a first-order (i.e. linear) polynomial is selected, and if the neighbourhood includes all the observations (i.e. $K = \mathrm{N}$ or $\alpha = 1$), this then results in the traditional linear regression. Thus, LOCFIT can be viewed as a superset. We used the software LOCFIT developed by Loader, which is available on-line (http://cm.bell-labs.com/cm/ms/departments/sia/project/locfit/index.html).

There are several nonparametric approaches to estimating the function f locally, such as kernel-based (Bowman and Azzalini 1997), splines, local polynomials (Owosina 1992; Rajagopalan and Lall, 1998; Loader, 1999). Owosina (1992) performed an extensive comparison of a number of regression methods, both parametric and nonparametric, on a variety of synthetic and real data sets. He found that the nonparametric methods handily outperform parametric alternatives. All of the nonparametric methods perform similarly, but LOCFIT is easy to implement; hence, we adopted it in this paper.

LOCFIT has been used for several hydroclimate applications (Lall, 1995): for spatial interpolation of precipitation (Rajagopalan and Lall, 1998); salinity modelling (Prairie, 2002; Prairie *et al*., 2005); stream flow modelling (Prairie, 2002; Prairie *et al*., 2004); stream flow forecasting (Grantz, 2003); and flood frequency estimation (Apipattanavis *et al*., 2005).

Variants of LOCFIT also provide an attractive alternative to ensemble generation. For example, the K neighbours of an estimation point x_i identified can be resampled (i.e. bootstrapped) with a weight function that gives more weight to the nearest neighbour and less to the farthest, thus generating ensembles. Lall and Sharma (1996) developed this approach and used it for stream flow simulation. Later, Rajagopalan and Lall (1999) and Yates *et al*. (2003) extended it for stochastic daily weather generation and Souza and Lall (2003) applied it for stream flow forecasting.

4.3. LOCFIT with resampled residuals (modified K-NN)

Frequently the errors e_t are not normally distributed. To address this issue a modification to LOCFIT was developed by Praire (2002). Prairie *et al*. (2004,2005) applied this for stream flow and salinity modelling. Later, Grantz (2003) demonstrated the use of this approach for stream flow forecasting on the Truckee–Carson basin in Nevada, USA.

Prairie (2002) referred to this as the 'modified K-NN', and we do the same here. The modification is described below.

Suppose an ensemble is required for a new value of the predictor x_{new}, and suppose that the polynomial order p and the size of the neighbourhood K have been obtained using GCV or other objective criteria. The steps in the modification are as follows:

1. Identify K nearest neighbours to x_{new} and fit a polynomial of order p. The fitted polynomial provides the estimate of the dependent variable at all the neighbours and, consequently, the residuals.
2. The fitted polynomial from step (1) is used to estimate the mean value Y_{new}. (This step is just the LOCFIT process described in the previous section.)
3. Now select one of the K neighbours of x_{new}, say x_i and select the corresponding residual e_i (already obtained from step 1); this is now added to the mean estimate $Y_{\text{new}} + e_i$, thus obtaining one of the ensemble members. The selection of one of the neighbours is done using a weight function

$$W(j) = \frac{1}{j\sum_{i=1}^{K}\frac{1}{i}} \tag{4}$$

As can be seen, this weight function gives more weight to the nearest neighbour and less to the farthest neighbours. Repeat step 3 several times, resulting in an ensemble.

The number of neighbours for fitting the local polynomial can be different from the neighbours used to resample the residuals (e.g. Prairie, 2002). In this study we have kept both the same. In the modification described above, if the number of observations N is small, then the resampled residuals (step 3 above) provide very limited variety in the ensembles, and this is the main disadvantage.

5. MODEL EVALUATION

The models are verified in a cross-validated mode, i.e. the data (rainfall and the predictors) for a given year are dropped out and the model(s) based on the rest of the data is (are) applied to generate an ensemble forecast for the dropped year. This is repeated for all the years for the 1980–2000 period. Apart from visual inspection, the ensembles are evaluated on three criteria:

1. Correlation between the observed value and the median of the ensemble forecast. This is much like evaluating the mean forecast that would come from a standard linear regression model.
2. Likelihood function (LLH; Rajagopalan *et al.*, 2002). This evaluates the skill of the model in capturing the probability density function (PDF).
3. Rank probability skill score (RPSS; Wilks, 1995). This evaluates the skill of the model in capturing the categorical probabilities, i.e. the PDF.

The likelihood function (LLH) is applied to measure the skill of forecasting models. Its process is to categorize predicted values into three categories; below normal, normal and above normal. The ensemble forecasts falling into these three categories are compared with historical data and then used to develop a skill score. The likelihood skill score for any given year of forecast is defined as

$$\mathrm{LLH} = \frac{\prod_{t=1}^{N} \hat{P}_{j,t}}{\prod_{t=1}^{N} P_{cj,t}} \tag{5}$$

where N is the number of years to be forecasted, j is the category of the observed value in year t, $\hat{P}_{j,t}$ is the forecast probability for category j in year t, and $P_{cj,t}$ is the climatological probability for category j in year t. Here, we divided the rainfall into three categories at the 33rd and 66th percentiles, so the probabilities of each of the categories are 1/3 and N is the length of data. The LLH values vary from zero to the number of categories (i.e. three in this study). A score of zero indicates lack of skill, a score greater than one indicates that the forecasts have skill in excess of the climatological forecast and a score of three indicates a perfect forecast.

The ranked probability skill score (RPSS) is also applied to quantify the skill of forecasting models. This method evaluates the probability of ensemble forecasts falling into many categories (i.e. below average, average and above average in this study) and compared with historical data. The RPSS score for any given year is defined as

$$\mathrm{RPS}(p, d) = \frac{1}{R-1}\left[\left(\sum_{i=1}^{R} P_i - \sum_{i=1}^{R} d_i\right)^2\right] \tag{6}$$

for R mutually exclusive and collectively exhaustive categories (in this case we have three categories, so $R = 3$). The vector d (d_1, d_2, ..., d_R) represents the observation vector such that $d_R = 1$ if the observation fell in category R or $d_R = 0$ otherwise. The RPSS is then calculated as (e.g. Toth, 2002; Wilks, 1995)

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS(forecast)}}{\mathrm{RPS(climatology)}} \tag{7}$$

RPSS scores vary from $+1$ to $-\infty$ (i.e. perfect skill to bad skill). Scores above zero indicate improvement over the climatological forecast.

For the LOCFIT and modified K-NN methods, owing to the small sample size we used polynomial of order ($p = 1$, i.e. local linear fit). However, the neighbourhood size was obtained objectively using the GCV criteria.

6. RESULTS

From the set of predictors (based on SST, SAT, SLP fields and ENSO indices) identified in Section 5, the optimal subset was found by the combination that gave the best forecast skill. Several formal methods are available for subset selection, such as stepwise regression, cross-validation metrics, etc. Since the number of significant predictors is small, in our case almost all combinations were tried out to find the optimal predictor set. For summer monsoon rainfall, the best set of predictors was found to be the one based on SLP and SST described in Section 3.3. The land temperatures (SAT) did not seem to improve the skill much. Forecasts were issued at the beginning of each month starting on 1 April for each year using all three methods. The predictors are the average values from the preceding season (i.e. preceding 3 months), except for forecasts issued on 1 July and 1 August, where the SST predictor of MAM and the SLP predictor of the preceding season are used, as this combination gave the best skill. Thus, the 1 July forecast is based on the SST predictor of MAM and the SLP predictor of April–June, and the 1 August forecast is based on the SST predictor of MAM and the SLP predictor of MJJ.

The skill of the forecasts is evaluated using the three skill measures described in Section 5. Model skill is also compared during high (wet) and low (dry) years. Threshold exceedance probabilities during the extreme years are estimated and the PDFs of ensembles of a few representative years are also presented.

The skill scores are shown in Figure 4. It can be seen that the skill increases significantly as the forecast lead time decreases for all the methods. This is intuitive and consistent with expectations. The linear regression and LOCFIT show similar skill on all three measures. This indicates that, for the most part, the relationship between the predictors and the rainfall is linear and that linear regression seems appropriate. The modified K-NN is comparable in performance as the lead time decreases, but its performance is weak early on. This, we believe, is due to the small sample size of residuals used in resampling. Given that we only have 21 data points (the data used for forecasting is for the period 1980–2000), K tends to be of the order of 7–8. With a small K, coupled with the fact that at long leads the relationship between predictors and rainfall is not as strong; consequently, there is less variety in the ensembles and a bias if the predictors are not very useful, which leads to poor skill scores.

Notice the significant skill from 1 May onwards, which provides a 2 month lead time that can be very useful for resources planning and management. This useful long-lead skill, regardless of the method, is quite impressive.

In order to compare the performance of these models in extreme years, Figure 5(a) and (b) shows the skill scores for the high (wet) and low (dry) rainfall years defined in Singhrattna (2003). Interestingly, the nonparametric models (LOCFIT and modified K-NN) seem to show a slight improvement over linear regression for forecasts starting 1 May in both the wet and dry years and generally for all the skill measures. This could be explained by the fact that a subtle nonlinear relationship exists between the predictors and the rainfall at the extremes; hence, there is some advantage in using nonparametric methods. Furthermore, notice that the skill in wet years is much more than that in the dry years.

PDFs of the ensemble forecasts (solid line) made on 1 August during selected wet and dry years from the three models are presented along with the *climatological PDF* (dashed line, which is estimated from the entire historical record) and the observed values (dotted line) in Figures 6 and 7 respectively. For the wet years the modified K-NN (Figure 6(a)) shows the ensembles to be shifted to the right of the climatological PDF. For the low years (especially, 1984 and 1994; Figure 7(b)) it can be seen that the LOCFIT method does a good job of shifting the forecast ensemble PDF to the left of the climatological PDF relative to the linear regression.

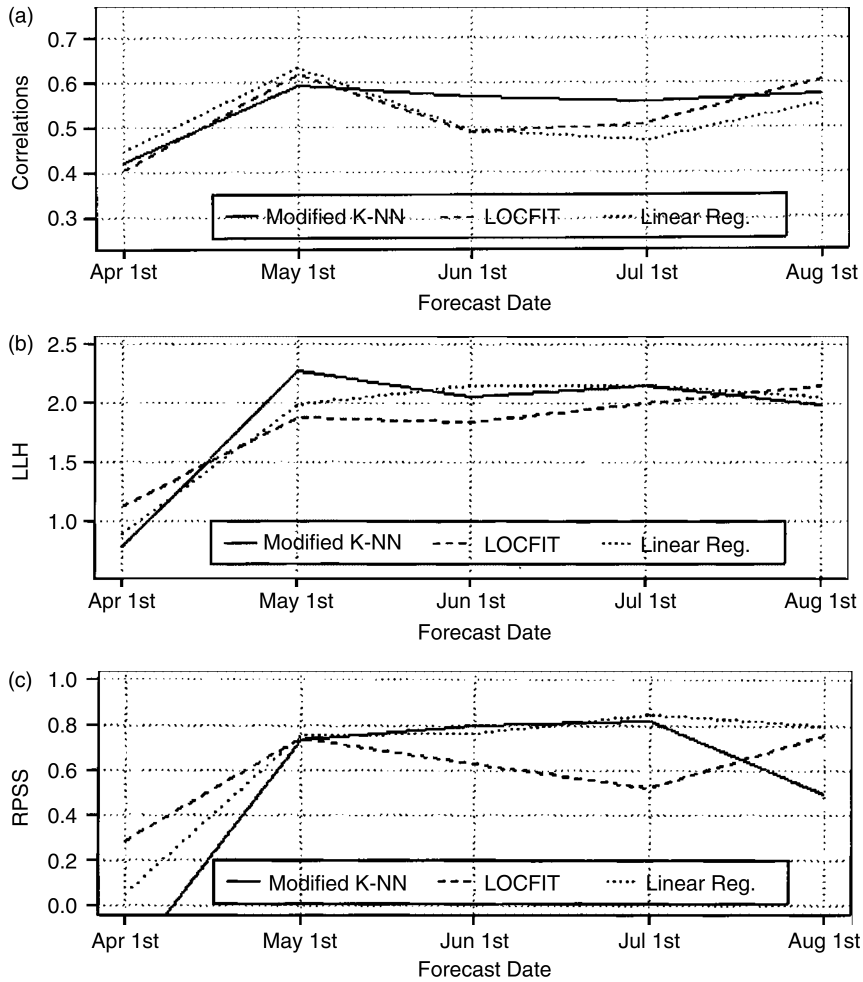

Figure 4. Cross-validated skill scores for Thailand summer rainfall forecasts issued on the first of each month from April through to August using the three ensemble forecast methods. The skill measures, i.e. correlation, LLH and RPSS, are shown in the three plots

Even though the observed values are not in the middle of the ensemble PDFs (as we would like them to be), this still can provide useful information and skill in terms of threshold exceedance probabilities, which is one of the key variables for making planning decisions. We chose 700 mm (the 90th percentile of the data) as a surrogate for wet (or flood) conditions and 400 mm (the 10th percentile of the data) for dry conditions. From the PDFs of the ensembles forecast on 1 May, the exceedance probabilities are computed for the selected wet and dry years and shown in Table II. For the wet years the climatological exceedance probability is 0.10, whereas the ensembles in all the years except 1995 indicate a very high probability of exceedance of this threshold, thus indicating a wet condition. This information, provided on 1 May, 3 months ahead of the summer monsoon season, could be very helpful in flood emergency response planning and management. For the dry years the models show a small non-exceedance probability of the lower threshold (400 mm) when a higher probability of non-exceedance is expected. This is consistent with the fact that the models have a low skill in dry years, especially with the 1 May forecasts (Figure 5(b)). However, we found the skill in these exceedance probabilities to be higher for 1 June, 1 July and 1 August forecasts. It can be seen that the nonparametric models, in general, show a slight improvement upon the linear regression model. Similar estimates were obtained from forecasts issued in other months.

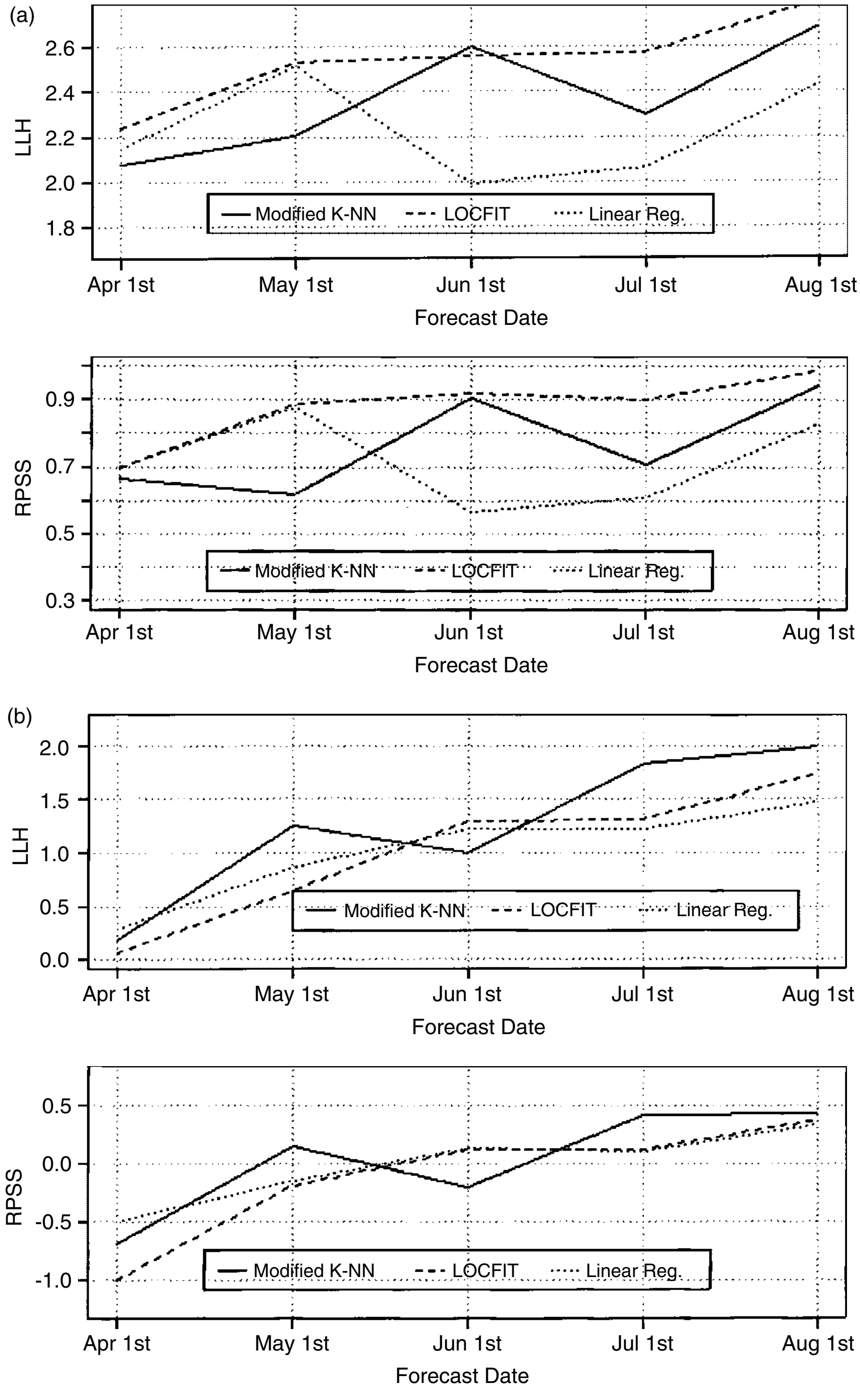

Figure 5. Same as Figure 4, but for (a) wet years and (b) dry years. Correlation measure is not shown owing to small sample size

The threshold exceedance probabilities can be used effectively to plan annual and seasonal reservoir, emergency response preparedness, floodplain management, cropping strategies, conservation measures, etc. Furthermore, they can also be used as a surrogate for wetness or dryness and provide probabilistic information on flooding potential, landslides, etc. and develop optimal response strategies. Lastly, the

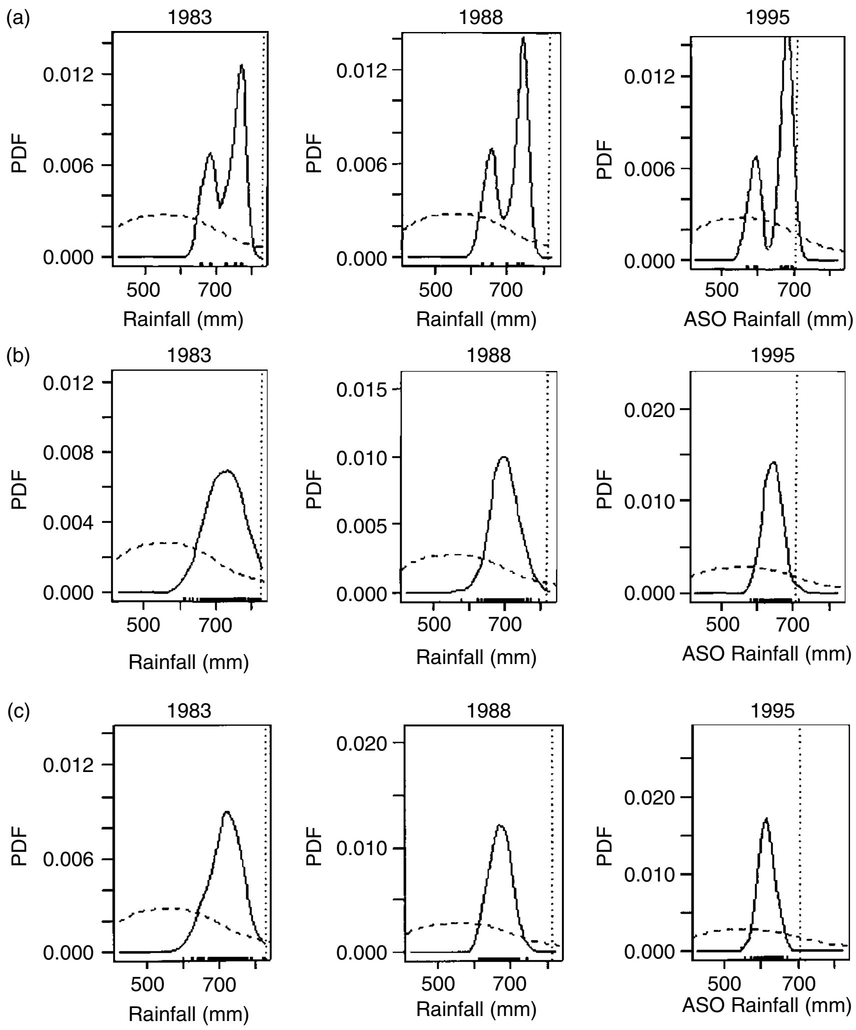

Figure 6. PDF of ensemble forecasts (solid line) and the climatological PDF (dotted line) for three selected wet years (1983, 1988 and 1995) from the three methods: (a) modified K-NN; (b) LOCFIT; (c) linear regression

ensembles of rainfall can be used to drive a water balance model and generate ensembles of stream flows. The forecasts will provide a useful and powerful tool to water managers in long-term planning that is currently lacking.

7. SUMMARY

Predictors from large-scale ocean, atmosphere and land variables that have strong correlations with Thailand summer monsoon have been identified. The predictors are consistent in terms of their physical mechanistic links to the monsoon. The predictors indicate that rainfall is predictable one to two seasons in advance. Interestingly, the predictors are related to the monsoon rainfall only during the post-1980 period, when the monsoon rainfall is correlated with ENSO, as seen in Singhrattna (2003). This suggests the tantalizing possibility that the ENSO relationship could be modulating the predictability, similar to what is seen in the case of the Indian monsoon (Krishna Kumar *et al.*, 1995, 1999b). The nonstationarity aspect of the relationship

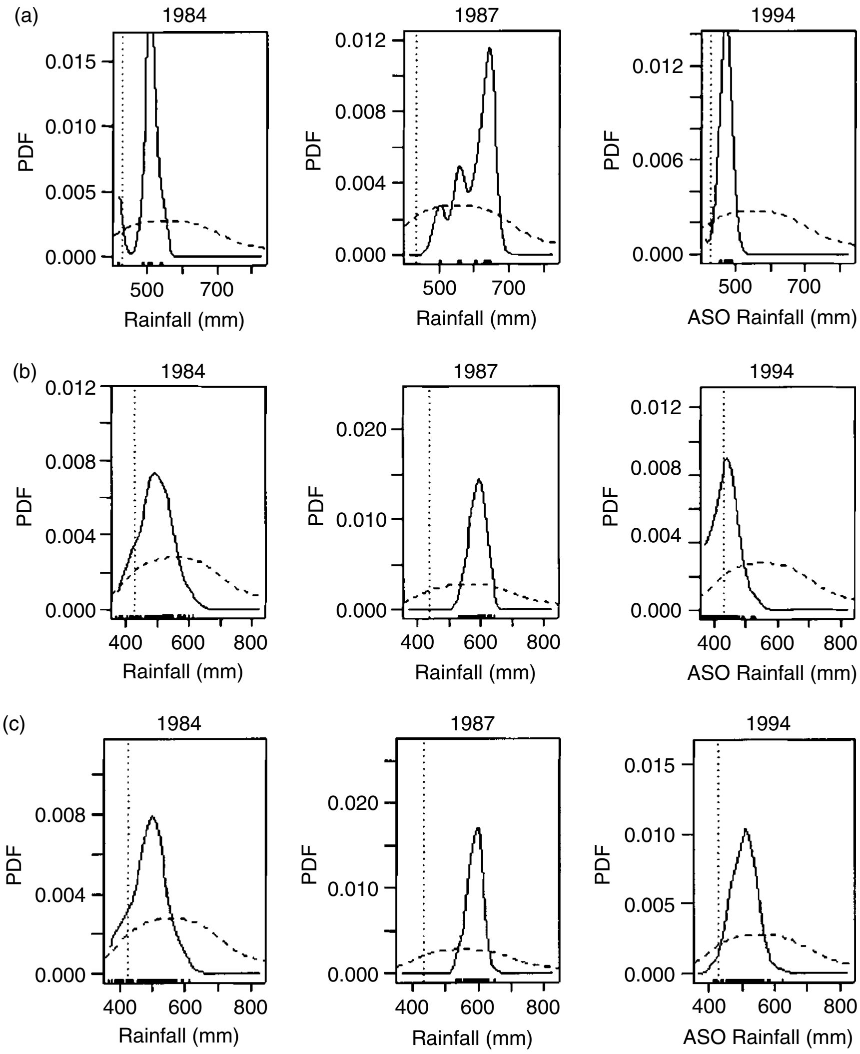

Figure 7. Same as Figure 6, but for selected dry years (1984, 1987 and 1994)

between the predictors and Thailand summer rainfall means caution must be exercised, in that the relationships have to be tested periodically and new predictors identified if necessary.

Two modelling approaches for ensemble forecasts of Thailand summer monsoon are offered: (1) a traditional linear regression (parametric) and (2) an adapted nonparametric method based on local polynomials. Both the models exhibit significant skill at 2–5 months' lead time. The nonparametric method seems to show improved skill in the extreme years, especially in wet years.

The proposed models for forecasting Thailand summer rainfall make a significant contribution, as no official forecast models exist to our knowledge. This has tremendous implications for water management, early warning and preparedness, and also for resources planning in general. Further testing and improvements of these models are required.

ACKNOWLEDGMENTS

NS is grateful to the Thailand Government for the graduate student fellowship and to the Thailand Public Works Department for their continued support. Travel support from IUGG to NS to present preliminary results

Table II. Exceedance probabilities for selected wet years and non-exceedance probabilities for selected dry years

Year	Probability (%)			
	Climatology	Modified K-NN	LOCFIT	Linear regression
1983	10.0	81.0	73.5	71.3
1988	10.0	39.9	54.7	33.6
1995	10.0	3.1	4.6	1.1
Dry years				
1984	10.0	1.0	1	1
1987	10.0	2.3	3.7	9
1994	10.0	1.0	1.5	1

from this work at the IUGG General Assembly 2003 in Sapporo, Japan, is very much appreciated. Funding for the KKK via a CIRES Visiting Fellowship is also gratefully acknowledged. We thank the GEWEX Asian Monsoon Experiment (GAME) for the precipitation data and for also providing a wealth of information on the monsoon efforts in Asia. We are grateful to two anonymous reviewers, whose comments greatly improved the manuscript. Finally, we would like to thank Somkiat Apipattanavis and Tom Chase for useful discussions.

REFERENCES

Apipattanavis S, Rajagopalan B, Lall U. 2005. Local polynomial technique for flood frequency analysis. *Journal of Hydrology* submitted for publication.

Bowman AW, Azzalini A. 1997. *Applied Smoothing Techniques for Data Analysis*. Clarendon Press: Oxford.

Bras RL, Iturbe IR. 1985. *Random Functions and Hydrology*. Addison-Wesley: Massachusetts.

Clark MP, Hay LE. 2004. Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *Journal of Hydrometeorology* **5**(1): 15–32.

Cordery I, McCall MA. 2000. A model for forecasting drought from teleconnections. *Water Resources Research* **36**(3): 763–768.

Fein JS, Stephens P (eds). 1987. *Monsoons*. Wiley–Interscience: New York.

Grantz K. 2003. Using large-scale climate information to forecast seasonal streamflows in the Truckee and Carson rivers. Master of Science thesis, University of Colorado, Boulder, CO.

Hastenrath S. 1987. On the prediction of Indian summer rainfall anomalies. *Journal of Climate and Applied Meteorology* **26**: 847–857.

Hastenrath S. 1988. Prediciton of Indian monsoon rainfall: further exploration. *Bulletin of the American Meteorological Society* **69**: 819–825.

Helsel DR, Hirsch RM. 1995. *Statistical Methods in Water Resources*. Elsevier Science: Amsterdam.

Jha R, Herath S, Musiake K. 1997. Development of IIS distributed hydrological model and its application in Chao Phraya river basin Thailand. *Annual Journal of Hydraulic Engineering, JSCE* **41**: 227–232.

Jha R, Herath S, Musiake K. 1998. Application of IIS distributed hydrological model in Nakon Sawan catchment Thailand. *Annual Journal of Hydraulic Engineering, JSCE* **42**: 145–150.

Ju J, Slingo JM. 1995. The Asian summer monsoon and ENSO *Quarterly Journal of the Royal Meteorological Society* **122**: 1133–1168.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**: 437–471.

Kanae S, Oki T, Musiake K. 2001. Impact of deforestation on regional precipitation over the IndoChina peninsula. *Journal of Hydrometeorology* **2**: 51–70.

Kripalani RH, Kulkarni A. 1997. Rainfall variability over Southeast Asia — connections with Indian monsoon and ENSO extremes: new perspectives. *International Journal of Climatology* **17**: 1155–1168.

Kripalani RH, Kulkarni A. 1998. The relationship between some large-scale atmospheric parameters and rainfall over Southeast Asia: a comparison with features over India. *Theoretical and Applied Climatology* **59**: 1–11.

Kripalani RH, Kulkarni A. 2001. Monsoon rainfall variations and teleconnections over South and East Asia. *International Journal of Climatology* **21**: 603–616.

Krishna Kumar K, Soman MK, Rupa Kumar K. 1995. Seasonal forecasting of Indian summer monsoon rainfall: a review. *Weather* **50**(12): 449–467.

Krishna Kumar K, Rajagopalan B, Cane MA. 1999a. On the weakening relationship between Indian monsoon and ENSO. *Science* **284**: 2156–2159.

Krishna Kumar K, Kleeman R, Cane MA, Rajagopalan B. 1999b. Epochal changes in the Indian monsoon–ENSO precursors. *Geophysical Research Letters* **26**(1): 75–78.

Lall U. 1995. Recent advances in nonparametric function estimation. *Reviews of Geophysics* **33**: 1093–1102.

Lall U, Sharma A. 1996. A nearest neighbour bootstrap for resampling hydrologic time series. *Water Resources Research* **32**(3): 679–693.

Loader C. 1999. *Local Regression Likelihood*. Springer: New York.
Lui Z, Valdes JB, Entekhabi D. 1998. Merging and error analysis of regional hydrometeorological anomaly forecasts conditioned on climate precursors. *Water Resources Research* **34**(8): 1959–1969.
McCabe G, Dettinger MD. 2002. Primary modes and predictability of year-to-year snowpack variations in the western United States from teleconnections with Pacific Ocean climate. *Journal of Hydrometeorology* **3**: 13–25.
Meehl GA, Arblaster JM. 1998. The Asian–Australian monsoon and El Niño–southern oscillation in the NCAR climate system model. *Journal of Climate* **11**: 1356–1385.
Miyakoda K, Kinter III JL, Yang S. 2003. The role of ENSO in the South Asian monsoon and pre-monsoon signals over the Tibetan plateau. *Journal of Meteorological Society of Japan* **81**(5): 1015–1039.
Ntale HK, Gan TY, Mwale D. 2003. Prediction of East African seasonal rainfall using canonical correlation analysis. *Journal of Climate* **16**(12): 2105–2112.
Owosina A. 1992. Methods for assessing the space and time variability of ground water data. *MS thesis*, Utah State University.
Pant GB, Parthasarathy B. 1981. Some aspects of an association between the southern oscillation and Indian summer monsoon. *Arch. Meteorol. Geophys. Biokl., Sr. B.* **29**: 245–251.
Piechota TC, Chiew FHS, Dracup JA, McMahon TA. 2001. Development of an exceedence probability streamflow forecast. *ASCE Journal of Hydrologic Engineering* **6**(1): 20–28.
Prairie JR. 2002. Long-term salinity prediction with uncertainty analysis: application for Colorado River above Glenwood Springs. MS thesis, University of Colorado at Boulder.
Prairie J, Rajagopalan B, Fulp T, Zagona E. 2005. Statistical nonparametric model for natural salt estimation. *ASCE Journal of Environmental Engineering* **131**(1): 130–138.
Prairie J, Rajagopalan B, Fulp T, Zagona E. 2005. A modified *K*-NN model for generating stochastic natural streamflows. *Journal of Hydrologic Engineerng* under revision.
Rajagopalan B, Lall U. 1998. Locally weighted polynimial estimation of spatial precipitation. *Journal of Geographic Information and Decision Analysis* **2**(3): 48–57.
Rajagopalan B, Lall U. 1999. A nearest neighbour bootstrap resampling scheme for resampling daily precipitation and other weather variables. *Water Resources Research* **35**(10): 3089–3101.
Rajagopalan B, Lall U, Zebiak S. 2002. Optimal categorical climate forecasts through multiple GCM ensemble combination and regularization. *Monthly Weather Review* **130**: 1792–1811.
Rajeevan M. 2001. Prediction of Indian summer monsoon: status, problems and prospects. *Current Science* **81**: 1451–1457.
Rasmusson EM, Carpenter TH. 1983. The relationship between the eastern Pacific sea surface temperature and rainfall over India and Sri Lanka. *Monthly Weather Review* **111**: 354–384.
Sahai AK, Grimm AM, Satyan V. 2000. All India summer monsoon rainfall prediction using an artificial neural network. *Climate Dynamics* **16**: 291–302.
Saji NH, Goswami BN, Vinayachandran PN, Yamagata T. 1999. A dipole mode in the tropical Indian Ocean. *Nature* **401**: 360–363.
Salas JD. 1985. Analysis and modeling of hydrologic time series. In *Handbook of Hydrology*, Mandment DR (ed.). McGraw-Hill: New York; 19.1–19.72.
Shabbar A, Barnston GA. 1996. Skill of seasonal climate forecasts in Canada using canonical correlation analysis. *Monthly Weather Review* **124**: 2370–2385.
Shukla J, Mooley DA. 1987. Empirical prediction of the summer monsoon rainfall over India. *Monthly Weather Review* **115**: 695–703.
Singhrattna N. 2003. Interannual and interdecadal variability of Thailand summer monsoon: diagnostic and forecast. Master of Science thesis, University of Colorado, Boulder.
Singhrattna N, Rajagopalan B, Krishna Kumar K, Clark M. In press. Interannual and interdecadal variability of Thailand summer monsoon. *Journal of Climate*.
Souza FA, Lall U. 2003. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: applications of a multivariate, semi-parametric algorithm. *Water Resources Research* **39**: 1307–1320.
Thapliyal V. 1981. ARIMA model for long-range prediction of monsoon rainfall in peninsula India. India Meteorological Department Monograph Climatology 12/81.
Torrence C, Webster P. 1999. Interdecadal changes in ENSO–monsoon system. *Journal of Climate* **12**(8): 2679–2690.
Toth Z. 2002. Assessing the value of probabilistic forecasts from a scientific perspective, validation of probabilistic forecasts. In *Predictability Seminar*, ECMWF, 9–13 September.
Walker GT. 1924. Correlation in seasonal variations of weather, IV, a further study of world weather. *Memoris of the Indian Meteorological Department* **24**: 275–332.
Wang B, Wu R, Li T. 2003. Atmosphere–warm ocean interaction and its impacts on Asian–Australian monsoon variation. *Journal of Climate* **16**: 1195–1211.
Webster P, Magana VO, Palmer TN, Shukla J, Tomas TA, Yanai M, Yasnari T. 1998. Monsoons: processes, predictability, and the prospects for prediction. *Journal of Geophysical Research* **103**(C7): 14 451–14 510.
Wilks DS. 1995. *Statistical Methods in the Atmospheric Science: An Introduction*. Academic Press: San Diego.
Yates D, Gangopadhyay S, Rajagopalan B, Strzepek K. 2003. A nearest neighbour bootstrap technique for generating regional climate scenarios for integrated assessments. *Water Resources Research* **39**(7): 1199, SWC 7-1–7-15.
Yevjevich VM. 1972. *Stochastic Processes in Hydrology*. Water Resources Publications: Fort Collins, CO.

