# D10 Predictor selection methods

**Colin Harpham and Rob Wilby, Kings College London, November 2003**

## Introduction

Three methods were employed to identify downscaling predictors that have the most significant influence on the predictand, namely, stepwise multiple linear regression (SWLR), a ranking procedure and a Genetic Algorithm. A wet day threshold of 1 mm/day was generally employed, however, other thresholds (5, 10, 15 and 20 mm/day) were also used to test their affect on predictor selection by the stepwise multiple linear regression. Two UK regions were used for this comparison the NWE and SEE and to compare the models the observed/modelled daily time series was analysed using the STARDEX diagnostic tool to produce indices of extreme precipitation.

## The model

For all the comparison experiments a Radial Basis Function (RBF) artificial neural network (Broomhead and Lowe, 1988; Moody and Darken, 1989) was employed to model the daily area average time series for the two regions. The RBF network consists typically of two layers, where the hidden layer nodes contain prototype vectors (or basis centres), which are in effect hidden layer weights. The distance between the input and the prototype vector determines the activation level of the hidden layer with the nonlinearity provided by a basis function. The activation function in the output layer can be nonlinear, however, training is considerably faster if an ordinary linear weighted sum of these activations are performed, and this approach was consequently adopted. Mathematically the output from the final layer node(s) $y_k$ (for the $k^{th}$ output node) is expressed by the following:

$$y_k(\mathbf{x}) = \sum_{j=1}^{M} w_{kj}\phi_j(\mathbf{x}) + w_{k0}$$

Where $\mathbf{x}$ is the $d$-dimensional input vector with elements $x_i$, and $w_{kj}$ are the output layer weights together with the bias $w_{k0}$, and $M$ is the number of hidden layer nodes or basis centres. The basis function $\phi_j(\mathbf{x})$ provides the nonlinearity and the final layer weights are calculated by a matrix inversion technique using singular value decomposition. To make comparisons the modelled and observed daily time series were run though the diagnostic tool to calculate the indices.

## Stepwise linear regression procedure

Series of area-average precipitation occurrence were prepared for the various wet-day exceedance thresholds (see above). Stepwise linear regression was then applied using all predictor variables (see Table 1) for coincident, lag-1 and lag+1 daily time steps. The selection process was halted once the improvement in the correlation between observed and fitted series was less than 1%. Independent validation data (1979-1993) were used to verify the skill of predictors selected from the calibration period (1961-1978; 1994-2000). Wet day amounts modelling was undertaken using the same thresholds, noting that increasing wet-day thresholds led to progressively smaller training sets and 10mm/day proved to be the upper limit to achieve worthwhile results. For example the NWE calibration set had only 67 days greater than the 20 mm threshold equating to just 0.9% of the data.

## Ranking procedure

Stepwise regression predictor selection is biased towards reproducing *mean* wet-day amounts. In order to weight the predictor variable selection towards *extreme* events the following approach was trialed. For each region, area-average amounts were ranked in descending order from largest to smallest daily precipitation totals. The average (normalised) values for all predictors were then calculated for the top 100 events. In this way, predictors resulting in the most extreme events were ranked. The statistical significance of the weights was tested using a bootstrap method in which 10,000 means were calculated for 100 randomly selected (with replacement) predictor values. Ranked means lying outside the 2.5 and 97.5 percentiles of the bootstrapped distribution were deemed to be significantly different from chance.

## Genetic algorithm

The pioneering work of Holland (1975) illustrated how the Darwinian evolution process could be applied, in the form of an algorithm, to solve a wide variety of problems. Due to the biological motivation, this highly parallel adaptive system is now called the genetic algorithm (GA). The GA has a population of individuals competing against each other in relation to a measure of fitness, with some individuals reproducing, others dying, and new individuals arising through combination and mutation. In this instance, the GA uses the trained RBF network as a fitness function so that the modelled/observed error could be minimised to evolve an optimal subset of predictor variables, which had been encoded into each member of the population.

## Results

**Tables 2** and **3** show the leading predictor variables identified by stepwise linear regression (without seasonal stratification) for NWE and SEE respectively using a 1mm/day threshold. **Tables 2** and **3** also report the ten most heavily weighted predictors identified by the ranking procedure for each region. The results highlight the gain of employing forward lagged predictors, which compensate for the mismatch between the NCEP day and the precipitation day. In contrast, backward lagged predictors were noteworthy by their absence, possibly reflecting the use of grid boxes to the west of the target regions (i.e., spatio-temporal substitution). Overall, the secondary (airflow) variables tended to dominate over the moisture variables for both selection methods and regions.

Vorticity at the surface (ZSUR) and at 850 hPa (Z850) were the most frequently selected variables for precipitation occurrence and amounts in both regions, consistent with previous analyses for the UK (Conway et al., 1996). Specific humidity at 500 hPa (Q500) was second for both methods in NWE and for the SWLR set in SEE. Beyond this, there was little consistency amongst the variables chosen by the two methods. Ranking ranked mean sea level pressure (MSLP) in third position for the NWE and fifth for the SWE. However, this variable was absent from the SWLR set. Relative humidity at 500 hPa (R500) and 850 hPa (R850) were notable omissions by both methods given their importance in previous studies (e.g., Beckman and Buishand, 2001).

Predictor variables from the Wales grid box tend to dominate NWE amounts models; whereas the balance was more even between Wales and Ireland for occurrence models. Similarly for SEE the SE grid box dominates. This highlights the need to include both propinquitous and remote predictors in the precipitation downscaling (Wilby and Wigley, 2000; Brinkmann, 2002).

The wet-day threshold was also found to influence the predictors selected by SWLR (not shown). For example, a threshold of 5 mm/day yielded only two predictors in common with the 1 mm/day threshold model for NWE for both the occurrence (Q500WA, U850WA) and amounts (Q500IR, ZSURWA). However, there was greater overlap between the SWLR amounts model and the predictors selected by ranking (Q500IR, H850WA, ZSURWA, F500WA) suggesting some convergence between the methods at higher thresholds. This might be expected because the smallest wet-day total in the NWE (SEE) ranked set was 18.48 (17.74) mm, and the average daily total was 22.80 (23.43) mm.

In contrast, results for SEE showed more agreement between the 1mm/day and 5mm/day SWLR amount models (ZSURSE, Q500SE, FSURSE, QSURSE) and less with the ranking (only ZSURSE). Overall, models with predictors obtained for higher thresholds yielded less explained variance than the 1mm/day models in both NWE and SEE.

Correlation coefficients for the RBF/observed indices for NWE and SEE area-averages indicated that the SWLR had an advantage over the ranking procedure in the vast majority of cases although the gains were marginal in the case of NWE (**Figures 1** and **2**).

As an aside, the GA approach yielded far too many predictors for statistical regression applications despite efforts to constrain selections via penalty functions and, consequently, was pursued no further.

## Discussion and conclusions

Models of area-average and multi-site daily precipitation were compared for northwest (NWE) and southeast (SEE) England with particular emphasis on extreme events. It was found that stepwise linear regression (SWLR) was generally more effective than ranking for predictor variable selection when a Radial Basis Function (RBF) model of area average precipitation was used to assess predictor skill against data not used for calibration (1979-1993). However, the overall gain was marginal, although it was originally hoped that because the ranking procedure is not dependent on a linear predictor/predictand relationship and is directly targeted at extremes some improvement in the prediction of extremes would be obtained. Unfortunately this was not the case and may be due to the squared error of the SWLR process having a greater influence on extremes than originally supposed.

The importance of performing separate SWLR for precipitation occurrence and amounts series was shown by differences in the resulting predictor sets. The SWLR and ranking analyses further confirm the over-riding importance of vorticity and specific humidity as downscaling predictors for daily precipitation. The wet-day threshold amount was also found to affect predictor selection.

## References

Beckmann, B.R. and Buishand, T.A. 2001. KNMI contribution to the European project WRINCLE: downscaling relationships for precipitation for several European sites. Technical Report TR-230, KNMI, De Bilt.

Brinkmann, W.A.R. 2002. Local versus remote grid points in climate downscaling. *Climate Research*, **21**, 27–42.

Broomhead, D.S. and Lowe, D. 1988. Multivariable function interpolation and adaptive networks. *Complex Systems,* **2**, pp. 321-355.

Holland J. 1975. Adaptation in natural and artificial systems. Ann Arbor: University of Michigan Press.

Moody, J.E. and Darken C.J. 1989. Fast learning in networks of locally-tuned processing units. *Neural Computation,* **1**, 281-294.

Wilby, R.L. and Wigley, T.M.L. 2000. Precipitation predictors for downscaling: observed and General Circulation Model relationships. *International Journal of Climatology*, 20, 641–661.

**Table 1.** Candidate predictor variables available for each grid box.

| Predictor | Description |
|---|---|
| TEMP | Mean temperature at 2m |
| MSLP | Mean sea level pressure |
| H850 | 850 hPa geopotential height |
| H500 | 500 hPa geopotential height |
| USUR | Near surface westerly wind |
| U850 | Westerly wind at 850 hPa |
| U500 | Westerly wind at 500 hPa |
| VSUR | Near surface southerly wind |
| V850 | Southerly wind at 850 hPa |
| V500 | Southerly wind at 500 hPa |
| FSUR | Near surface wind strength |
| F850 | Wind strength at 850 hPa |
| F500 | Wind strength at 500 hPa |
| ZSUR | Near surface vorticity |
| Z850 | Vorticity at 850 hPa |
| Z500 | Vorticity at 500 hPa |
| DSUR | Near surface divergence |
| D850 | Divergence at 850 hPa |
| D500 | Divergence at 500 hPa |
| QSUR | Near surface specific humidity |
| Q850 | Specific humidity at 850 hPa |
| Q500 | Specific humidity at 500 hPa |
| RSUR | Near surface relative humidity |
| R850 | Relative humidity at 850 hPa |
| R500 | Relative humidity at 500 hPa |

**Table 2.** Selection of predictor variables for NWE precipitation using the sub-periods 1961-1978 and 1994-2000. The first four characters in each predictor name refer to Table 3, the last two characters specify the grid box (WA = Wales, IR = Ireland). The wet-day threshold was 1.0 mm/day.

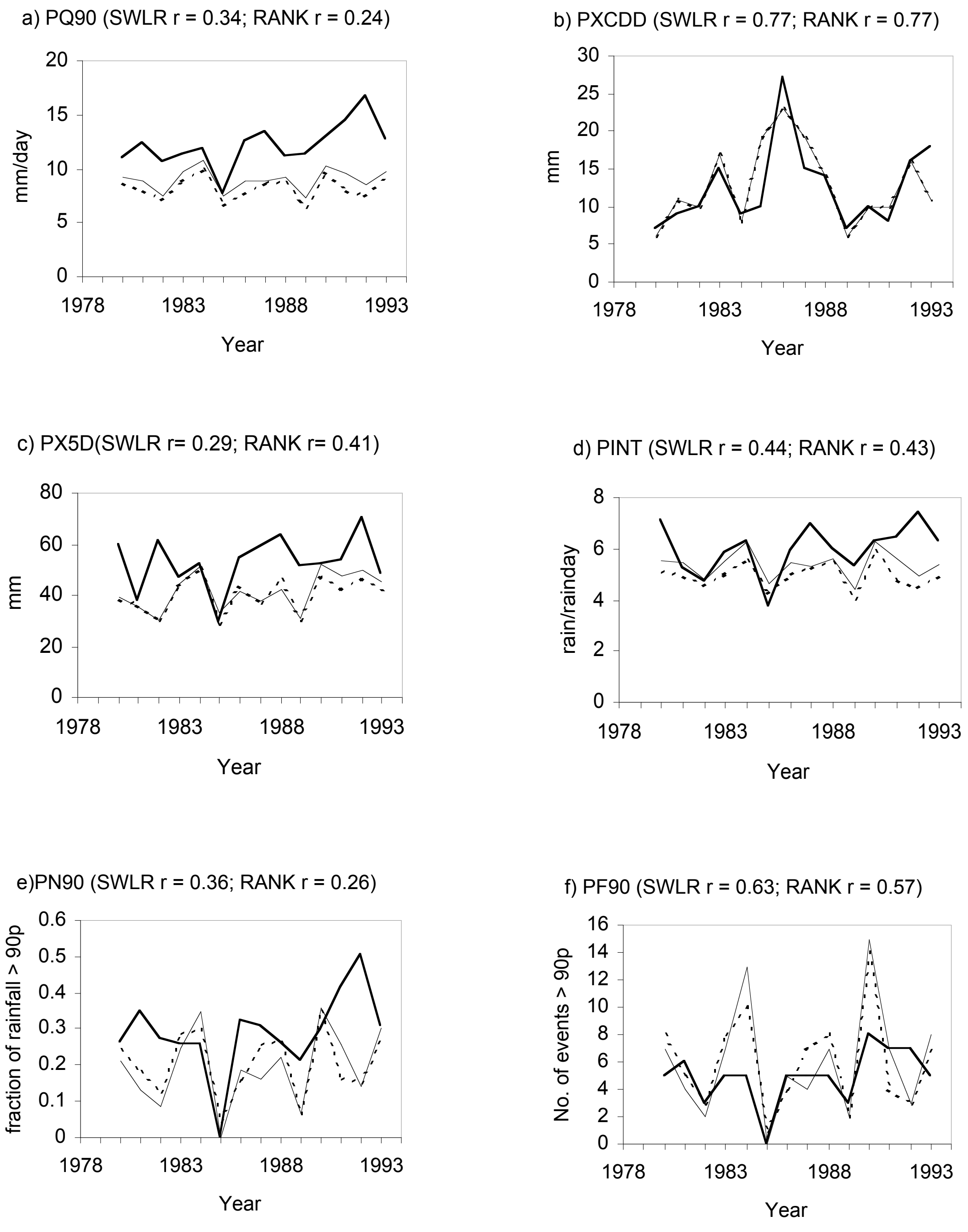
| **Step** | **Occurrence** | | | **Amounts** | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***Predictor (SWLR)*** | ***Lag*** | ***r*** | ***Predictor (SWLR)*** | ***Lag*** | ***r*** | ***Predictor (composited)*** | ***Lag*** | ***Weight*** |
| 1 | **U850WA** | **+1** | **0.49** | **ZSURWA** | **+1** | **0.30** | **ZSURWA** | **+1** | **1.38** |
| 2 | **Z850WA** | **+1** | **0.63** | **Q500WA** | **0** | **0.42** | **Q500WA** | **0** | **1.28** |
| 3 | **ZSURWA** | **0** | **0.65** | **F850WA** | **+1** | **0.49** | **MSLPWA** | **+1** | **1.16** |
| 4 | **RSURIR** | **0** | **0.67** | **ZSURWA** | **0** | **0.50** | **Q500IR** | **0** | **1.14** |
| 5 | **USURWA** | **0** | **0.68** | **U500WA** | **+1** | **0.52** | **Z850WA** | **+1** | **1.10** |
| 6 | F500WA | +1 | 0.68 | **Z500WA** | **+1** | **0.53** | **F500WA** | **+1** | **1.08** |
| 7 | Q500WA | 0 | 0.68 | **S500IR** | **-1** | **0.54** | **F850WA** | **+1** | **1.07** |
| 8 | H500IR | 0 | 0.69 | **ZSURIR** | **+1** | **0.55** | **U500WA** | **+1** | **1.01** |
| 9 | RSURWA | +1 | 0.69 | QSURIR | 0 | 0.55 | U850WA | +1 | 0.99 |
| 10 | Q850IR | 0 | 0.70 | TEMPWA | +1 | 0.56 | H850WA | +1 | 0.98 |

Note: **bold** type indicates the predictor variables used.

**Table 3.** As Table 2 but for SEE. Again, the last two characters of the predictor name specify the grid box (SW = South West England, SE = Southern England).

| **Step** | **Occurrence** | | | **Amounts** | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***Predictor (SWLR)*** | ***Lag*** | ***r*** | ***Predictor (SWLR)*** | ***Lag*** | ***r*** | ***Predictor (composited)*** | ***Lag*** | ***Weight*** |
| 1 | **ZSURSW** | **+1** | **0.53** | **ZSURSE** | **+1** | **0.36** | **ZSURSE** | **+1** | **1.87** |
| 2 | **MSLPSW** | **0** | **0.59** | **Q500SE** | **0** | **0.43** | **Z850SE** | **+1** | **1.66** |
| 3 | **MSLPSE** | **0** | **0.61** | **VSURSE** | **0** | **0.48** | **Z850SW** | **+1** | **1.54** |
| 4 | **RSURSE** | **+1** | **0.62** | **FSURSE** | **+1** | **0.50** | **ZSURSW** | **+1** | **1.51** |
| 5 | **Z500SW** | **+1** | **0.64** | **Z850SE** | **+1** | **0.51** | **MSLPSE** | **+1** | **1.39** |
| 6 | **ZSURSE** | **0** | **0.65** | RSURSE | +1 | 0.51 | MSLPSE | +1 | 1.33 |
| 7 | F500SE | +1 | 0.65 | DSURSE | 0 | 0.52 | H850SE | +1 | 1.19 |
| 8 | USURSW | 0 | 0.66 | F500SE | +1 | 0.52 | H850SW | +1 | 1.17 |
| 9 | ZSURSE | +1 | 0.66 | F850SE | 0 | 0.52 | Z500SW | +1 | 1.09 |
| 10 | Q500SE | 0 | 0.67 | QSURSE | 0 | 0.53 | Z500SE | +1 | 1.06 |

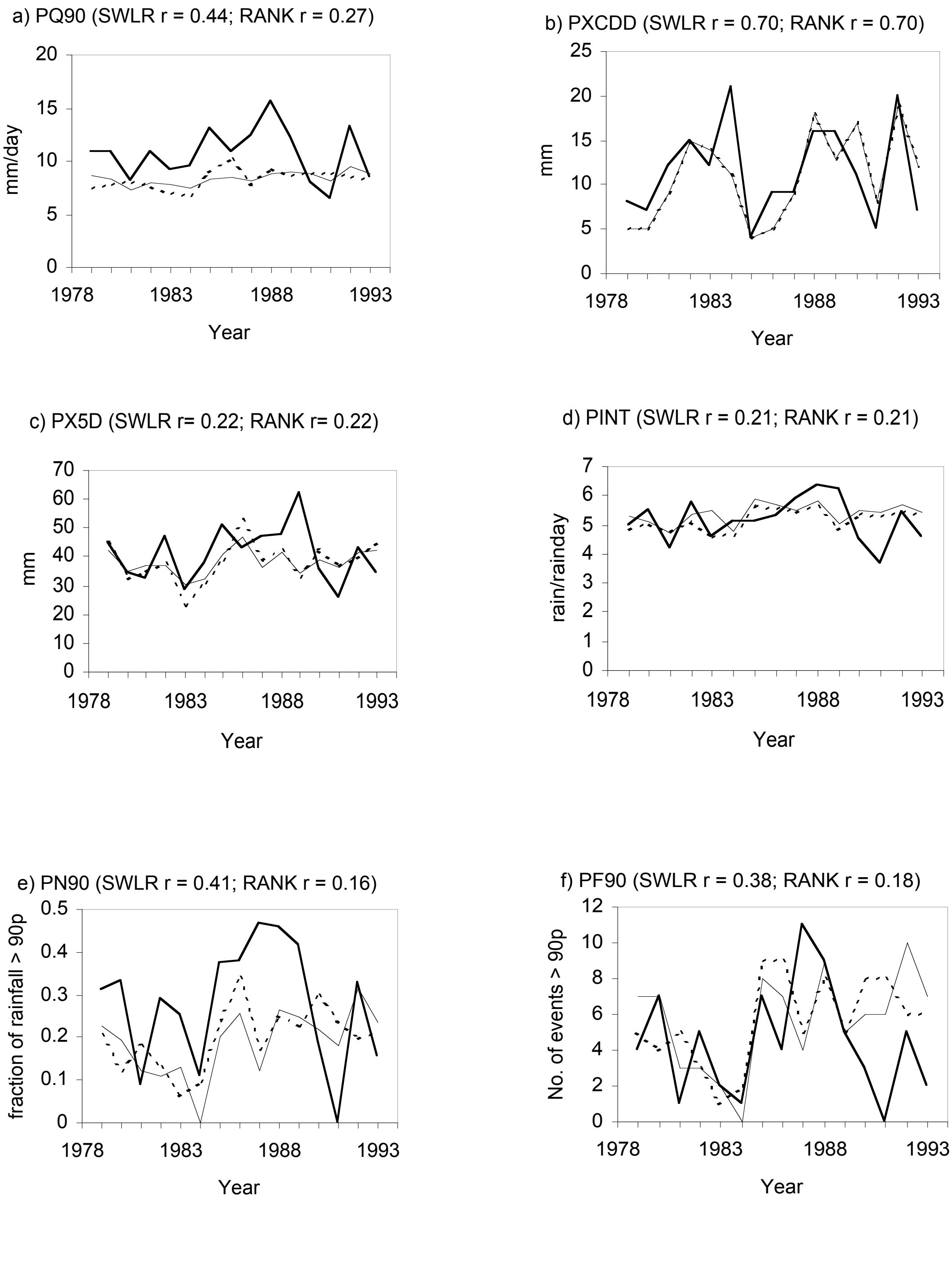
**Figures 1.** NWE comparisons of observed and modelled indices during DJF for the validation period 1979-1993.




## Legend

Observed - Thick line
RBF modelled (SWLR) - Thin line
RBF modelled (Ranked) – Dotted

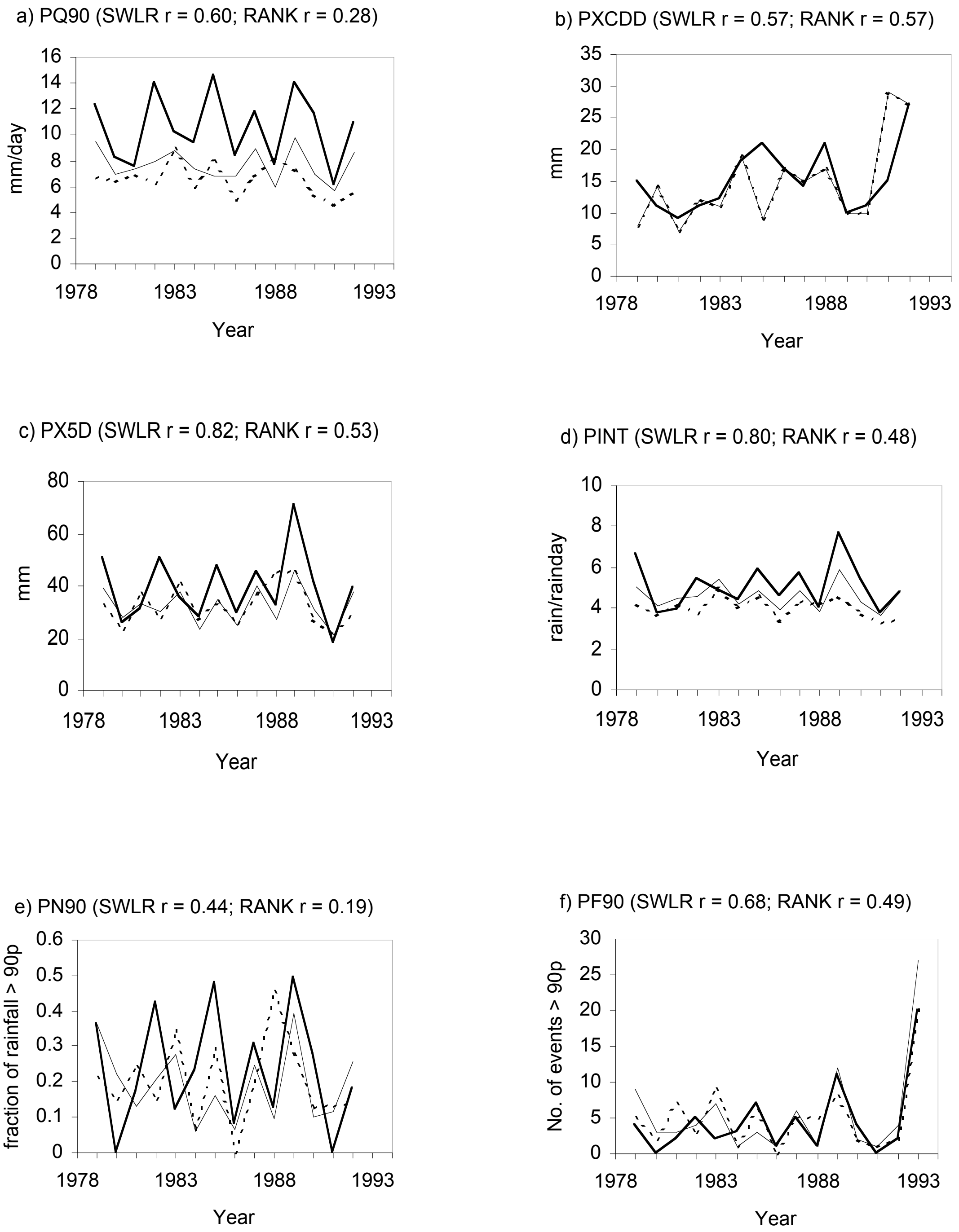
**Figures 2.** NWE comparisons of observed and modelled indices during JJA for the validation period 1979-1993.




## Legend

Observed - Thick line
RBF modelled (SWLR) - Thin line
RBF modelled (Ranked) – Dotted

**Figures 3** SEE comparisons of observed and modelled indices during DJF for the validation period 1979-1993.



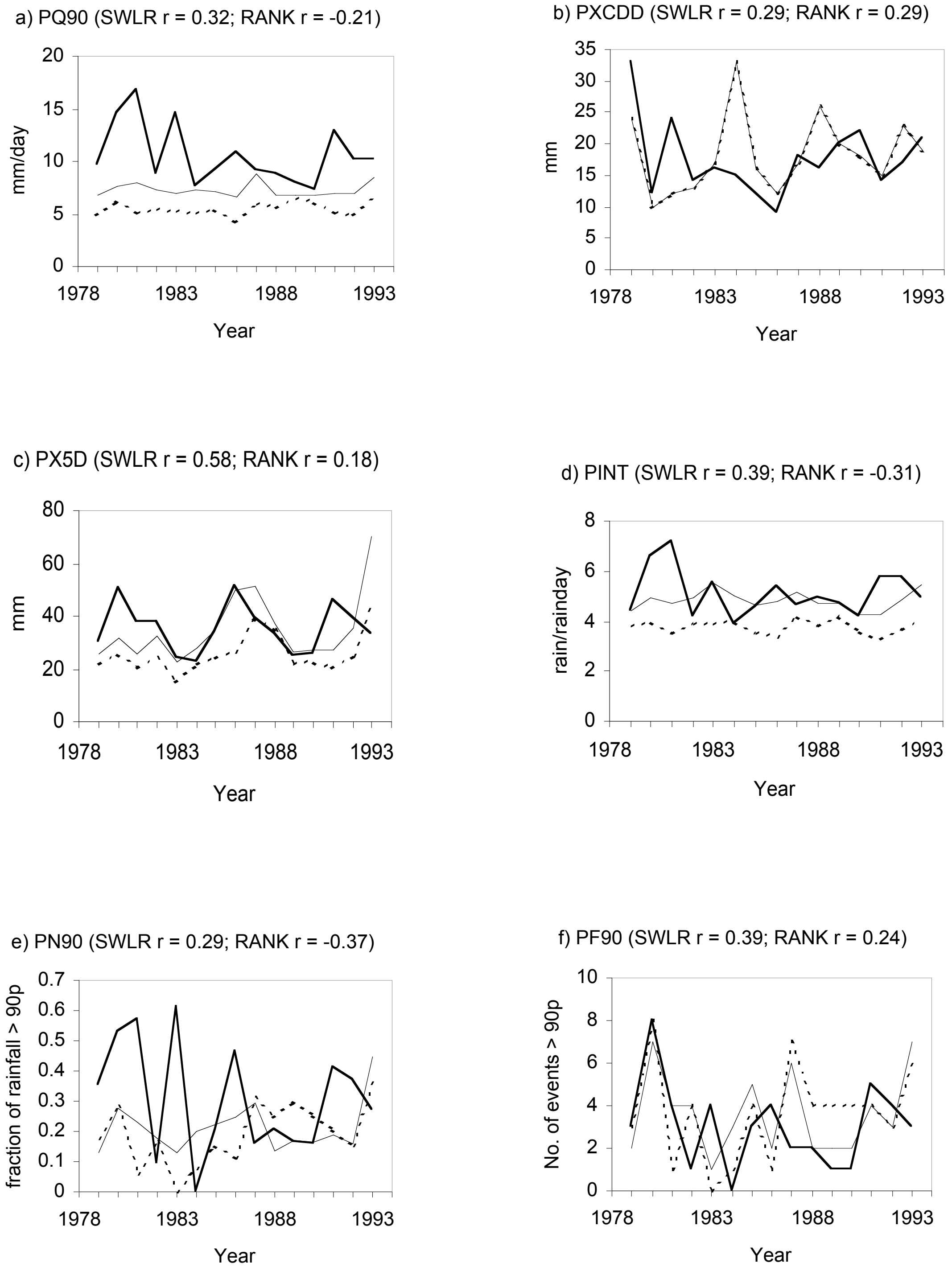

## Legend

Observed - Thick line

RBF modelled (SWLR) - Thin line

RBF modelled (Ranked) – Dotted

**Figures 4** SEE comparisons of observed and modelled indices during JJA for the validation period 1979-1993.




## Legend

Observed - Thick line
RBF modelled (SWLR) - Thin line
RBF modelled (Ranked) – Dotted