利用主成份迴歸分析探討區域單一場量與台北降雨機率的關係

林秀雯　　陳雲蘭
預報中心
中央氣象局

摘　　要

本文希望利用主成份迴歸（Principle Component Regression）方法，使用歐洲中期天氣預報中心模式十年冬季（1980至1989，10月至2月）的各場量資料（U、V、H、T...）與臺北站的雨量資料作迴歸相關預報。此法首先利用主成份分析法（Principle Component Analysis），對資料作主成份抽取，刪減訊息微弱的主成份，由於各主成份之間爲無相關，因此由各主成份與預報元的相關結果，可直接相加而得最後的迴歸關係。

初步結果顯示，在各層場變數測試中，以850百帕渦度場及U分量場的預報得分較高。而十年發展資料穩定測試發現前二年（1980及1981）的資料特性與其他年有差異。另外，由各分月測試情形，不同月份得分也有顯著的差異，原因可能與資料量限制、空間解析度、季節分類選取以及年際變化等因素有關，需作更多的測試研究。

一、前言

本局自民國82年元月開始對外發布降水機率預報，至今仍僅有氣候統計資料作爲預報的參考，尚未利用數值模式產品進一步發展較佳之統計客觀指引，如MOS（Model Output Statistics）及完全預報法模式PP（Perfect-Prog）。

近年來，本局在數值預報方面投入相當的努力，並且也有顯著的成就，因此我們希望能利用這些數值預報的資料用來作各種天氣要素的客觀機率預報。但不管採用MOS或是PP模式，都需要對過去的資料作迴歸分析，而如何針對預報元（Predictand）在這些龐大的氣象資料中選取合適的預報因子，作好迴歸分析，就成了另一個考慮的因素了。

爲求得一個較有效率的迴歸分析方法，我們引用主成份迴歸分析來測試模式的作業與預報成效。由於此法先利用主成份分析法，把資料中高度內相關的情形減至最低，並且去除幾近爲雜訊的主成份，因此可將資料作有效的精簡。另外，由於以主成份取代原來未處理過的變數因子，各主成份之間爲無相關的特性，使我們可直接將各主成份與預報元的相關結果相加，而得最後的迴歸關係，免去逐步迴歸複雜且冗長的步驟。

現階段我們先從單一場量與臺北降雨機率的關係著手，根據過去經驗，首先選用850百帕U場與V場等資料，分別與臺北降雨機率作相關測試。

二、使用資料

由於本局第一代區域模式產品僅僅累積了三年的資料，而第二代模式亦正在發展當中，因此目前即將開始發展之降水機率完全預報法模式，決定先以歐洲中期天氣預報中心（ECMWF）的模式包含十年冬季（1980～1990）各垂直分層場資料與本局臺北站觀測的降水資料來建立迴歸方程。

研究的空間範圍設定在東經100度至140度，北緯10度至40度，網格間距爲2.5度×2.5度，時間間隔則爲12小時，囊括00Z及12Z的資料，並利用任九年的資料測試預報穩定情形。

三、主成份迴歸分析法

因爲大氣複雜的特性，往往需靠多種角度的因子才能有較好的掌握。另外我們也知道想要找到適合來詮釋預報元的因子，需要作大量的測試。因此尋求一套複迴歸統計預報模式，使其能快速的選取多元變數，提昇工作效率，是我們採用主成份迴歸作此研究的原因。

1、主成份分析

主成份分析又稱經驗正交分析，乃是將可能有相互關係存在的多個變數，改成由所有變數組成的線性組合，產生新的座標。而此種爲所有變數綜合特性的新組合，即稱爲主成份。

定義 $S=\sum A_i X^T$

其中

S爲主成份

A爲權重係數矩陣

X爲資料矩陣

對資料陣列乘上一個權重，每個資料網格有其特定的權重值。接下來我們希望按照主成份所含訊息的多寡，依序分爲第一主成份、第二主成份...，由此可將訊息集中到前面少數個主成份，如此可用較少數的主成份來代表原有的多個變數，將變數數目予以減少。

因此，在定義主成份後，我們的要求是在所有的個案中，主成份的變異數 V_{ss}有最大值。另外，爲避免變異數的無限放大，有邊界條件$\sum a_i^2=1$的限制。如此，得以下方程組：

$$S=\sum A_i X^T$$
$$\text{Max}(\ V_{ss})$$
$$AA^T=I$$

如此可求得第一個主成份，而第二個主成份是去掉第一個主成份已說明的訊息後，再求具有最大變異數的分量。依此類推，即可將座標轉至主成份的方向。配合矩陣的觀念，我們可由解固有結構問題的方法來求解：

$$\text{由}V_{ss}=\frac{1}{M\sum SS^T}$$

$$=\frac{1}{M\sum[AX][AX]T}$$

$$=AV_{xx}A^T$$

其中

M爲樣本數

V_{xx}爲原網格座標資料的變異-互變異矩陣

利用V_{xx}來求解V_{ss}，解$(V_{xx}-\lambda I)A=0$固有結構問題

由V_{xx}矩陣解出的固有值就等於主成份的變異數，而固有向量就等於權重係數。再由定義主成份的公式代入原來的變數值，即可求得由各變數線性組合後的主成份$[S]$，其中向量S的元素組就稱爲主成份。

以上透過主成份分析，我們依然保留了所有的資料訊息，只不過將說明訊息的座標轉到變數相互獨立的主成份上。因爲各主成份按變異數大小排列的緣故，使得前面的主成份有較多的訊息，而後面的主成份訊息微弱，也就是說透過主成份分析，可將訊息集中，而可由前面少數幾個主成份來代表原本多數但可能含有大量重覆訊息的變數。

2、主成份迴歸

用主成份替代原來的變數再與預報元作迴歸，就是所謂的主成份迴歸方法了。迴歸的公式與傳統的複迴歸並沒有兩樣，不過由於預報因子間有相互正交的特性，所以在作法上，只需直接累加即可，因此去掉了逐步迴歸的問題，使求算過程簡化不少。

主成份迴歸的另一個好處是資料的精簡。如前所述，主成份分析可將資料訊息集中，而排列在後面的主成份則多爲雜訊，所含之訊息較小，因此只需取前面幾個主成份，就可包含整筆資料大部份的訊息。在不影響資料訊息的前提下，主成份數目的大量減少，當然就提高了計算的速度，並且也減小了資料儲存的空間。

我們可藉由圖（一）來說明主成份迴歸分析的骨架，及與其他迴歸方法的差異。其中Row data 代表原始的資料型態，而Processed data 則爲透過一些過濾工具，作過某種程度資料濾波處理後的資料。因此路徑B與路徑C都是直接將選定的變數（x）與預報元（Y）作迴歸分析，是屬於傳統的複迴歸統計方法。

如果尋路徑D以後就是主成份迴歸了，透過主成份分析，抽取解釋資料訊息達99% 的主成份，再以此作爲預報因子的方法。由於主成份分析可集中資料訊息，所以第一次的主成份分析可刪減訊息微弱的主成份。如果將主成份再作一次主成份分析，可再去掉訊息過少的主成份，只截取解釋到一定要求程度的主成份。因此路徑E是更爲有效率的主成份迴歸步驟，其實它同時也是製作多層多場變數與預報元作統計迴歸所必須的設計，亦即先分別對各層場作第一次的主成份分析，再分別標準化後作第二次的主成份分析，以提供作爲預報用的預報因子，此多層場的主成份分析是我們未來也想嘗試努力的目標。

另外，路徑A是直接由原始資料作一次主成份分析，去掉訊息小的主成份來作複迴歸的方式，這也是我們目前測試階段採用的方法。因此PC —Reg 仍爲一種架構在迴歸方法上的處理方式。只是以PC先行處理資料，如此雖不會直接提高預報的準確性，但可提高我們尋找合適預報因子的效率，間接幫助預報的品質。

四、測試結果

配合EC十年資料，我們利用1980～1990年降水機率的資料來測試PC—REG的統計預報成效。測試的方法主要是以十年任取一年當作預報年，由其他的九年當發展資料來檢驗預報年的預報成效。技術得分（skill score）的求算方法如下：

$$\text{skill score}=\frac{\sum|\hat{Y}_{mi}-Y_m|^2-\sum|\hat{Y}_i-Y_m|^2}{\sum|\hat{Y}_{mi}-Y_m|^2}\times 100\%$$

其中

$\hat{Y}_m$爲由氣候值作爲預報估計值

Y_m爲預報年之氣候平均值

$\hat{Y}$為主成份迴歸預報法之預報估計值

測試的項目有針對預報因子選取的分層分場測試，希望了解那一個層場的因子有較好的預報。而不同延遲相關的測試則在了解將二組時間序列作些微的時間位移後，可否得到更好的相關。空間範圍的測試可幫助我們明瞭影響臺灣地區降雨因子的範圍，最後作不同發展資料年限的測試以了解樣本資料穩定的情形。

1、空間範圍的測試

在對台北的降水機率作預報，我們主觀地先選取EC2.5° X2.5° 網格，一足夠大的範圍10° N～40° N、100° E～140° E來作測試工作，參考過去影響台北的天氣系統，我們認為此一範圍作完全預報統計模式已是綽綽有餘了。

但我們仍想進一步在此範圍內作進一步的測試，我們將空間範圍縮小、或改變解析度來測試技術得分對空間範圍劃分的敏感度。如圖(二)的四種網格選取，分別為全部選取的221點（二、a）、去掉外圍所剩的165點（二、b）、減低解析度的63點（二、c）以及縮小範圍的90點（二、d）。

我們利用850百帕速度場(V)場作為說明變數來作此測試，其得分情形如(圖三、a～d)。由其可知：空間範圍的減小與解析度的降低，均將使技術得分略為降低。

因此，我們接下來均以全部網格點221點來進行測試。

2、分層分場的測試

我們以EC七層的垂直分層及各物理變數資料作最佳說明變數測試，希望挑選與台北降水機率最有關聯的因子。

初步先利用現有已檢查過的U、V及導出量輻合（散）場及渦旋場來測試，發現在垂直分層500百帕以上的變數，對地面的降水預報技術偏低，而在物理變數的表現方面，以850百帕為例，結果如圖（四），可看出以渦度場及U分量的得分較高，V分量較低。

3、不同延遲相關的測試

雖然傳統的完全預報法是利用同時相關建立的模式，但就統計預報的角度來看，兩組資料的最佳關係可能也有出現在延遲的情形。因此我們也作了延遲相關的測試，希望看看說明變數與臺北降水在時間方向的關係（圖五）。

在此以850百帕渦度場為說明變數，分別與降水作Lag-2～Lag+2的測試，結果如圖（六）。由於我們的資料時間間隔為12小時，所以相當於作前後24小時延遲的測試。結果顯示仍以Lag0的得分為最佳，Lag-1比Lag+1表現稍好，而Lag±2以上的技術得分就明顯下降了。

4、不同發展資料年限的測試

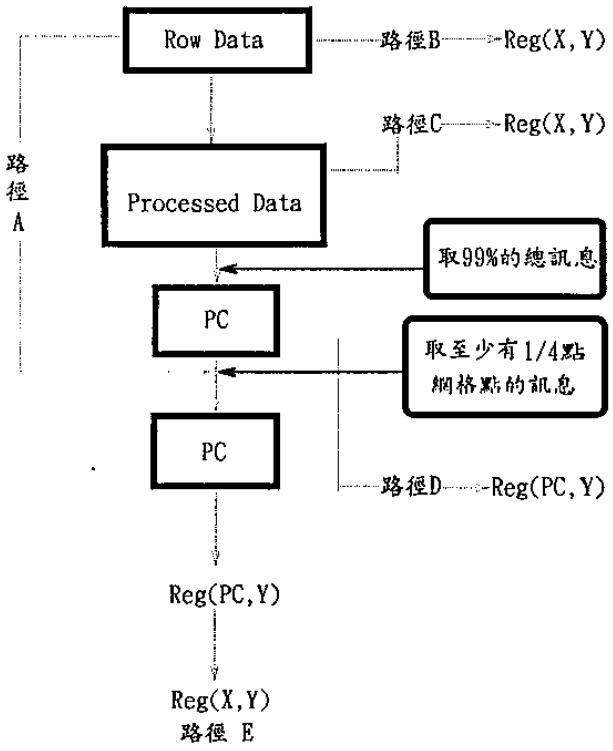
在利用任九年的發展資料作另外一年的預報技術測試時，我們發現各年的得分表現有明顯的差異，顯示這十年的資料特性並不是非常的穩定，因此我們根據得分較差的情形，抽取出1980-1981及1981-1982兩年冬季的資料，剩下的八年再以任七年作為發展資料，作另外一年的預報技術測試。同理，再抽取得分低的兩年，作任五年預報另外一年的情形，結果如圖（七）。由測試可知，去掉較難預報的年份，可使發展資料的統計特性較為一致，也使得預報模式較為穩定，因此由任五年預報另一年的得分最好。但是我們同時必須注意的是，太過短的發展資料使樣本數降低，將減少統計方法代表的可信度。

另外，我們也作了分月的測試，發現不同的月份有顯著的得分差異，這也提醒我們需對發展資料的季節劃分區間作一仔細的測試研究。

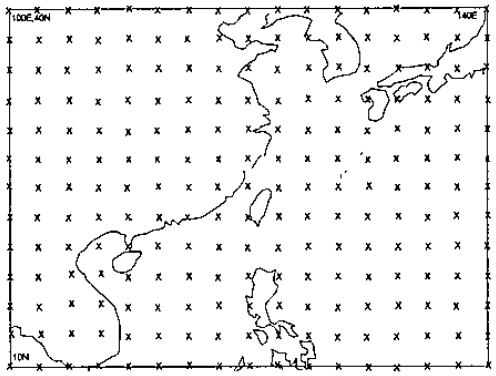
五、總結

選取合適的預報因子是想作好統計預報一定要考慮的問題，而這尋找的過程也代表需作大量的測試工作。若能有一套有效率的統計迴歸工具，不管在作業上或研究方面，都將可有很大的幫助。

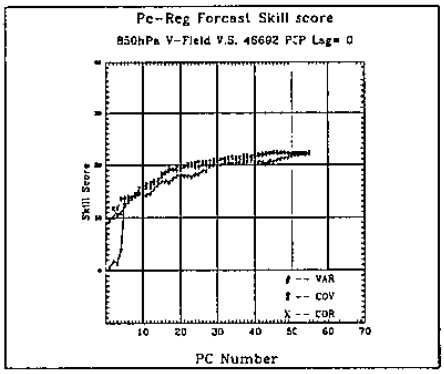
目前我們利用主成份迴歸作降水機率預報，現階段使用單一場量選取合適的空間、時間資料發展統計模式，其預報技術與氣候值比較已有不錯的成績。未來再進行多層場變數的組合，對預報元提供更多的說明訊息，眞正發揮主成份迴歸分析在統計預報的長處。



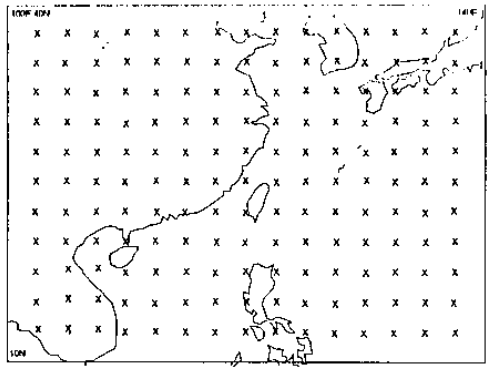
(圖一) 主成份迴歸法的簡單圖示



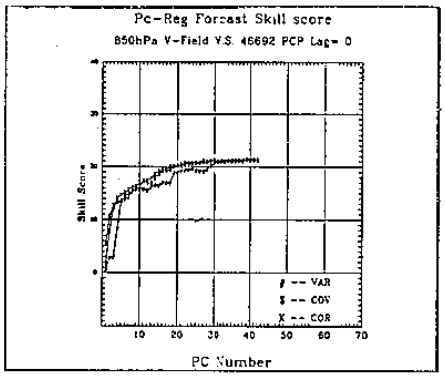
（圖二、a）空間範圍測試的選取，自10° N～40° N；100° E～140° E，每2.5° x2.5° 為一網格，以 "x" 表示之，如上圖所示共221個網格點。



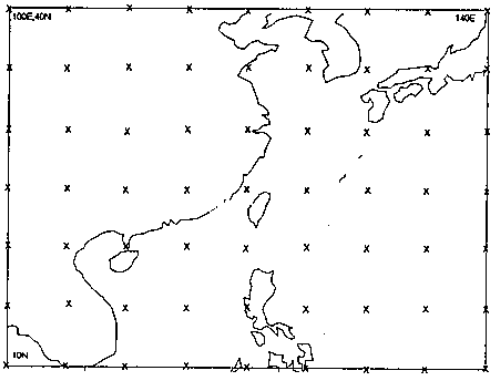
（圖三、a）利用850百帕V場作為說明變數測試得分情形，參考(圖二、a)，上圖為取用221點網格的測試結果。



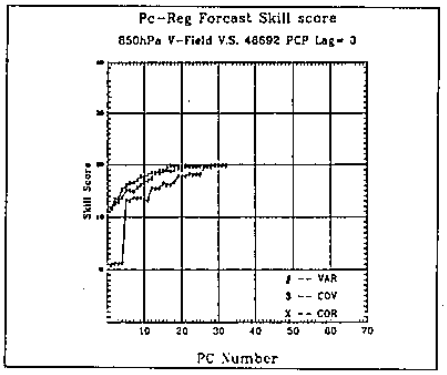
（圖二、b）同(圖二、a)，但共有165個網格點。



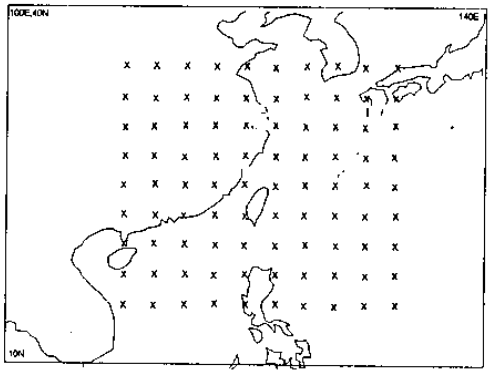
（圖三、b）同(圖三、a)但參考(圖二、b)，上圖為取用165點網格的測試結果。



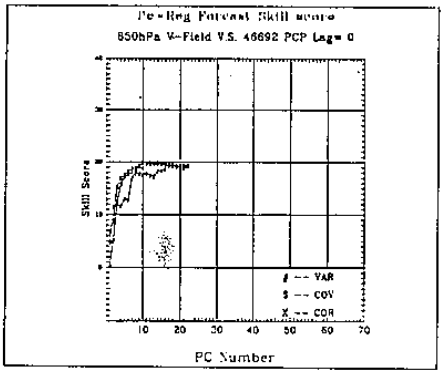
（圖二、c）同(圖二、a)，但共有63個網格點。



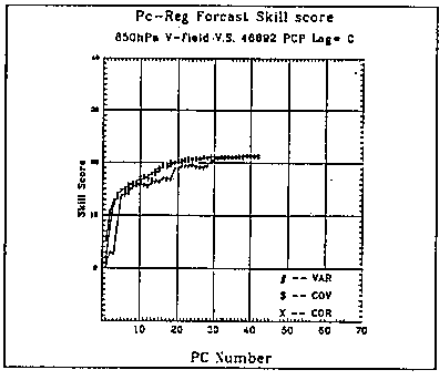
（圖三、c）同(圖三、a)但參考(圖二、c)，上圖為取用63點網格的測試結果。



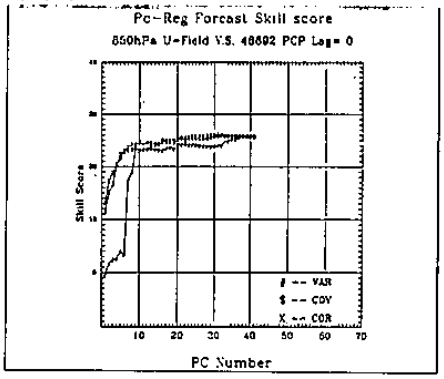
（圖二、d）同(圖二、a)，但共有90個網格點。



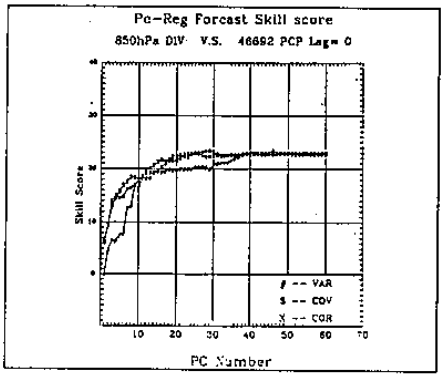
（圖三、d）同(圖三、a)但參考(圖二、d)，上圖為取用90點網格的測試結果。



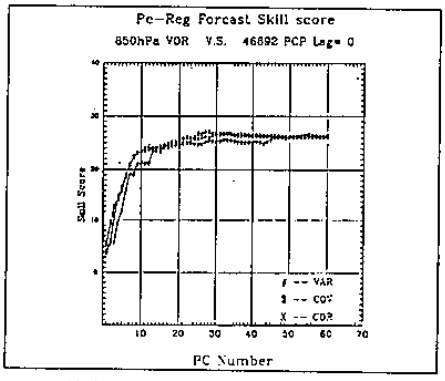
(圖四、a) 特定層(850hPa)的速度場(V)對台北降水機率預報技術測試。



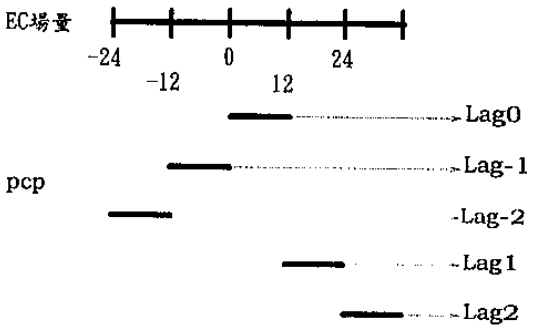
(圖四、b) 特定層(850hPa)的速度場(U)對台北降水機率預報技術測試。



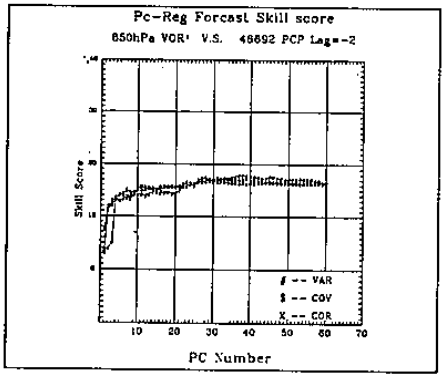
(圖四、c) 特定層(850hPa)的輻散場(DIV)對台北降水機機率預報技術測試。



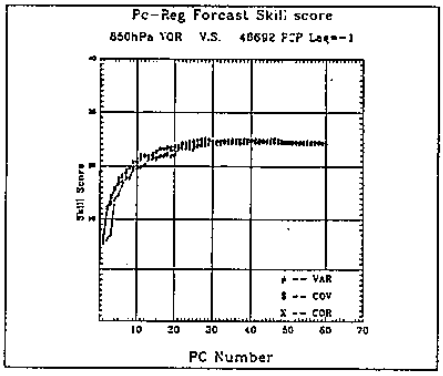
(圖四、d) 特定層(850hPa)的渦度場(VOR)對台北降水機機率預報技術測試。



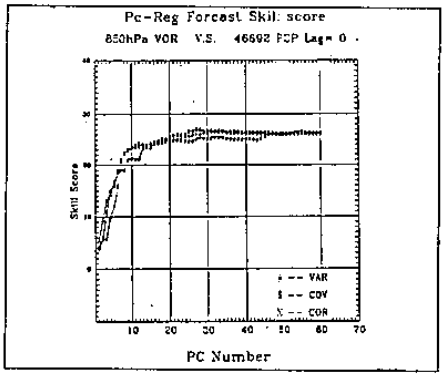
(圖五) 以EC場量分別與降水作Lag-2 ～Lag+2 的延遲相關測試之示意圖。



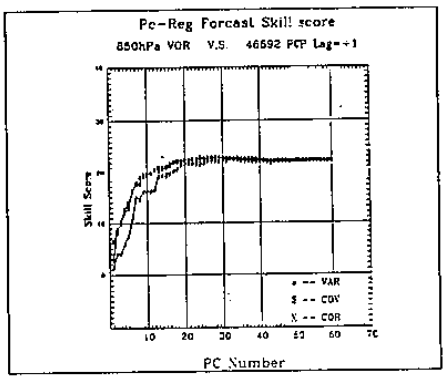
(圖六、a) 以850 百帕渦度場為說明變數，與降水作Lag-2 的延遲相關之測試。



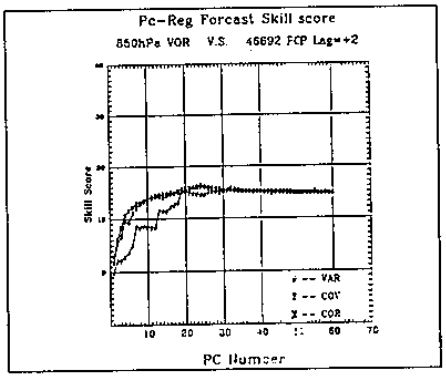
(圖六、b) 以850 百帕渦度場為說明變數，與降水作Lag-1 的延遲相關之測試。



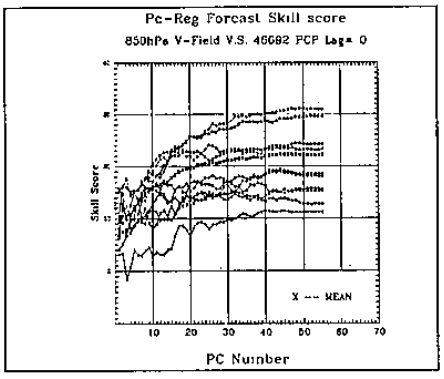
圖六、c) 以850 百帕渦度場為說明變數，與降水作Lag0 的延遲相關之測試。



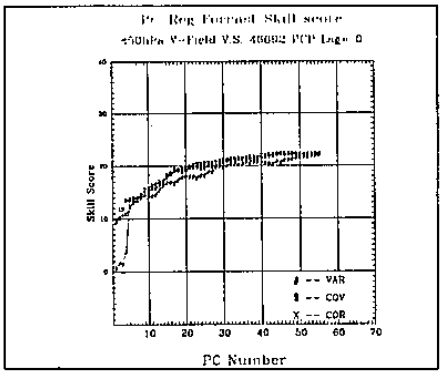
(圖六、d) 以850百帕渦度場為說明變數，與降水作Lag+1 的延遲相關之測試。



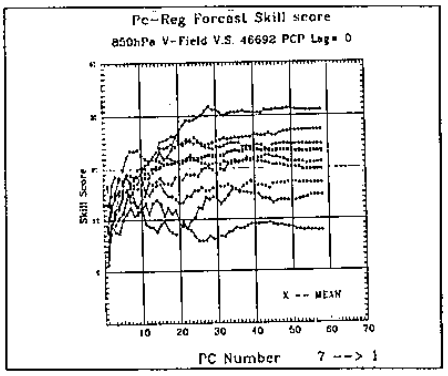
(圖六、e) 以850百帕渦度場為說明變數，與降水作Lag+2 的延遲相關之測試。



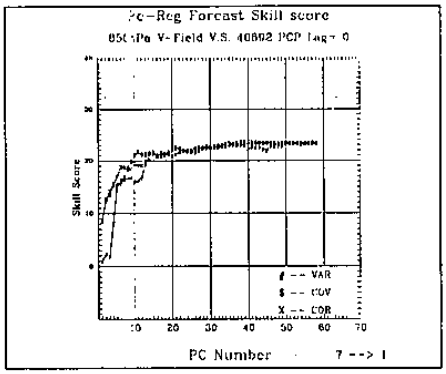
(圖七、a) 以任九年的發展資料進行另一年的預報測試。



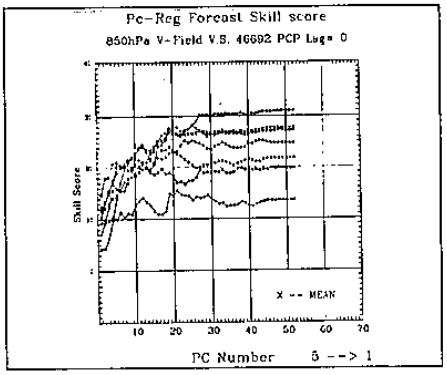
(圖七、b) 同(圖七、a)但為十組預報之平均得分。



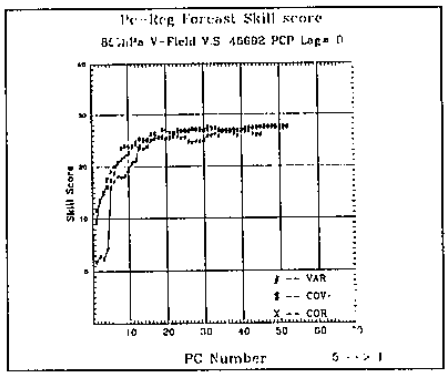
(圖七、c) 以任七年的發展資料進行另一年的預報測試。



(圖七、d).同(圖七、c)但為八組預報之平均得分。



(圖七、e) 以任五年的發展資料進行另一年的預報測試。



(圖七、f) 同(圖七、e)但為六組預報之平均得分。