應用隨機森林與支撐向量機於雨量與水位預報

林焜詳[1]、曾宏偉[2]、郭振民[3]、楊道昌[4]、游保杉[5]

摘要

本研究運用兩種機器學習演算法-支撐向量機與隨機森林，進行未來 1~3 小時雨量及水位預報。首先，以支撐向量機測試不同輸入因子於雨量預報之表現，結果顯示：輸入因子採用雨量、颱風以及天氣因子於雨量預報表現最佳。然後，進一步比較支撐向量機與隨機森林於雨量預報之表現，結果顯示：隨機森林之預報雨量較無時間延遲問題，且預報雨量表現穩定，不會隨著前置時間延長而有明顯變動；而支撐向量機於雨量預報之表現則會隨著前置時間變長而變差。最後，採用隨機森林建構水位預報模式，結果顯示：預報水位無延遲問題且可提供準確水位預報值。

關鍵詞：隨機森林、支撐向量機、雨量預報、水位預報

Application of Random Forest and Support Vector Machine in Rainfall and River Stage Forecasting

Kun-Xiang Lin[1], Hung-Wei Tseng[2], Chen-Min Kuo[3], Tao-Chang Yang[4], Pao-Shan Yu[5]

ABSTRACT

This study applied two machine learning methods, support vector machine (SVM) and random forest (RF), to predict rainfall and rive stage for one to three hours ahead. Firstly, various predictors were examined for finding a better model performance. The results show that using rainfall, typhoon factors and meteorological factors as input predictors can give best model performance. Then, the forecasting performance of SVM and RF were compared. The results show that the performance of SVM decreases while lead time increased. The performance of RF is relatively steady which does not vary with lead time and no time-lag problems are found in RF predictions. Finally, RF was selected for river stage forecasting. In general, RF can give reasonable and accurate estimates without time-lag problems.

Keywords: Random Forests、Support Vector Machine、rainfall forecasting、river stage forecasting

[1] 國立成功大學水利及海洋工程所 碩士生
[2] 國立成功大學水利及海洋工程所 博士後研究員
[3] 國立成功大學水利及海洋工程所 助理研究員
[4] 國立成功大學水利及海洋工程所 副研究教授
[5] 國立成功大學水利及海洋工程所 特聘教授兼工學院院長

一、緒論

洪水一直為國內重大天然災害之一，隨著氣候變遷，極端洪水事件更為嚴重，為減輕此種災害損失，洪水預警之重要性也更為明顯，因此發展一合適之洪水預報模式以提供決策者發佈預警之參考為一重要課題。

傳統水文方面較注重於逕流量之預報，即建立降雨-逕流模式以經由模式預測逕流量，再將預報逕流量轉換成預報水位以提供決策者參考。一般而言，降雨-逕流模式需利用歷史降雨事件之流量歷線進行率定與驗證，然而水位轉換成流量，或流量轉換為水位，均需仰賴流量率定曲線(rating curve)，而流量率定曲線則需藉由河川斷面及流速測量資料來製作，一般在高流量時較難以量測，在應用上常利用外插法予以延伸，造成利用降雨-逕流模式進行預報之不確定性。以洪水預報之觀點來說，在颱風、暴雨期間，實務單位皆採用河川水位作為預警之變量，若直接針對河川水位進行預報，則可免除流量與水位間的轉換及其不確定性因素，並可以較直接獲得預測之洪水是否達到警戒水位，因此採用河川水位作為預報變量較流量作為預報之變量更具實用性，故本研究之洪水預報直接採用水位作為預報變量。

對於水文模式而言，前述類型之傳統物理架構模式(physical based model)雖欲探討水文過程之物理機制，但時常面臨模式過於複雜，參數難由少數歷史觀測資料率定之問題，因此發展出較簡化的概念模式(conceptual model)，使其較易於應用，然而概念模式仍然需要有一定數量之歷史觀測資料來率定，且水文過程常具有非線性及隨機等特性。近年來，支撐向量機已成功運用於多種水文預報，如運用支撐向量機及模糊理論模式(陳憲宗，2005)成功進行洪水水位預報，此研究研究區域為蘭陽溪並收集 18 場颱風事件分為 12 場率定驗證資料與 6 場測試資料，藉由 10 摺交叉驗證(10 folds cross-validation)與網格搜尋法進行參數最佳化率定，建立未來 6 小時水位預報系統，結果顯示支撐向量機可有效地預測未來水位。除支撐向量機之外，另有一新興模式隨機森林(Random Forest)亦具有處理資料驅動模式能力，隨機森林由 Breiman 於 2001 年提出，其模式僅需決定兩個簡易參數且計算效率極高，並無過度擬合之疑慮，雖在水利領域之應用研究尚屬開端階段，但目前已成功應用在許多領域。

本研究參考隨機森林在河川水位即時預報之應用(郭家妏，2014)，並引入支撐向量機為理論依據，建立雨量預報及水位預報模式，雨量預報給予水位預報之預報輸入因子。然而，初步水位預報後，兩方法之水位預報均有時間延遲之問題，故額外使用真實雨量取代雨量預報作為輸入變量，結果大幅改善時間延遲之問題。因此，本研究參考颱風因子應用於颱風期間之雨量預報(林國峰，2013)，並蒐集更多集水區雨量站即時天氣資訊，期望能建立更精準的即時雨量預報，並增進水位預報之準確性。

二、研究區域與資料概述

2.1 研究區域介紹

本研究區域為宜蘭河流域，位於台灣東北部，流域面積 149.06 km^2，主流長 17.25 km，涵蓋礁溪、員山、宜蘭及壯圍等鄉市，流域內以中山橋為界，上游之地表坡度大於 0.24，而主流坡度大於 0.038。宜蘭河發源宜蘭縣礁溪鄉與新北市烏來區界雪山山脈的大礁溪和小礁溪山，其發源標高約 1160 m，西起雪山山脈之大、小礁溪與大湖溪沖積扇，東至壯圍沙丘海岸，地勢西高東低，呈坡度下降，其源頭為五十溪山西峰，大湖溪為宜蘭河之支流，於員山鄉員山大橋附近匯合五十溪後，至下游再與大礁溪、小礁溪會合，於新城附近匯集成宜蘭河主流，並於宜蘭市北邊壯圍鄉附近匯入蘭陽溪。目前宜蘭河流域主流已整治完成，但仍有內外水淹水問題(下水道排水問題、河口頂托效應)，且其流經宜蘭市區(農田、魚塭)，災害危害度較高，而流域所設置之水文測站資料較為齊全，因此選為本研究之研究區域。

2.2 水文資料概述

本研究主要目標為建立宜蘭河流域之即時河川雨量與水位預報模式，首先針對流域內現有水文測站蒐集相關資料。本研究不考量河口漲潮影響，目標水位站選為不受感潮影響之中山橋(西門橋)水位站，雨量站僅考量中山橋以上資料充足之雨量站，計有大礁溪、再連、雙連埤等共 3 站。以上資料蒐集包含時水位及時雨量資料，資料來源為台灣颱風洪水研究中心。除上游雨量站資料，另外蒐集中央氣象局提供之颱風特性因子以及宜蘭雨量站颱風期間之天氣因子，各雨量站位置如圖 1。

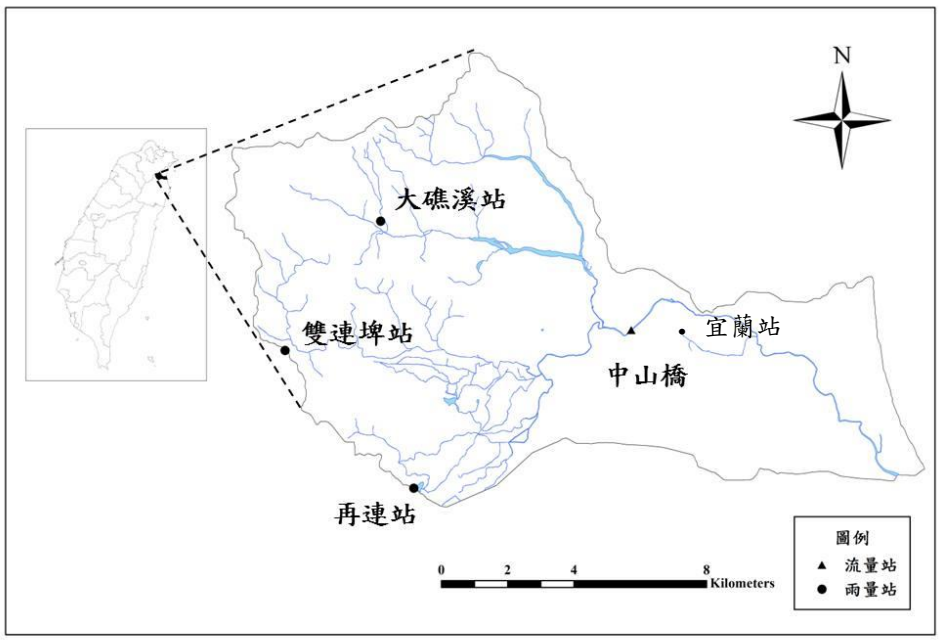

圖 1. 研究區域與水文測站

三、研究方法

3.1 隨機森林

隨機森林(Random Forests, RFs)是 Breiman and Culter (2001)所提出的整體式學習方法(ensemble learning)，並且把 Random Forests™ 作為此方法的商標。此方法因運用分類迴歸樹(Classification And Regression Tree, CART)，建構每一棵決策樹時使用隨機抽樣方式篩選資料樣本及資料變量，故有「隨機森林」之稱。

3.1.1 隨機性

隨機森林迴歸為一種整體式學習方法。整體式學習方法其主要概念是透過多次執行基礎學習演算法以建立多個基礎模型，並將每個基礎模型預測結果整理成較一致性之輸出決策，其結果會較傳統單一模型預測結果具更佳之績效表現。欲建構整體式學習系統其每個基礎模型必須具有足夠的準確度及基礎模型間必須具有多樣性，即每個基礎模型必須展現出不同之特性，才能發揮整體式學習之功效。

隨機森林之基礎學習演算法為未剪枝的分類迴歸樹，其建構整體式學習方法使用操控訓練樣本以及操控輸入特徵集，其在建立一棵迴歸樹時，分別隨機選取樣本資料與隨機選取輸入因子。隨機選取樣本資料係指隨機森林中樹的產生是利用 Breiman 於 1996 年所提出的拔靴法(bagging, bootstrap aggregating)進行資料重複抽樣，是採用均勻亂數的方式從 N 筆資料中選出 N 個子訓練集樣本，每次從原始訓練資料集中隨機選取子訓練集所用的資料，且選到的資料會放回訓練資料集中，故有些資料會重複被選取。N 次抽樣後每一個樣本沒被選到的機率大約為 0.368，如下式：

$$(1-\frac{1}{N})^N \approx e^{-1} = 0.368 \quad (1)$$

整體而言，抽樣結束後約有 1/3 的訓練資料沒有被選到，這些資料被稱為袋外資料(Out-Of-Bag data, OOB data)，可用來評估此棵迴歸樹的優劣及對輸入因子進行重要性評估。

隨機選取輸入因子係指隨機森林中每棵樹所考量的輸入因子，由使用者給定 m 值(m 值需小於 M 值)，採均勻亂數的方式從 M 個輸入因子中選出 m 個輸入因子出來，選出來的輸入因子不再放回樣本內，故為不重複選取。

以前述拔靴法選取樣本資料與隨機選取輸入因子方法完成一子訓練集之選，並以此子訓練集建構一棵迴歸樹，重複此步驟直至完成森林中樹的數量，亦即隨機森林模式完成，至此亦達成操控訓練樣本以及操控輸入特徵集之策略，使森林中之各迴歸樹間具有多樣性之特性。

森林建構完成後，欲使用隨機森林模式進行迴歸預測，其迴歸的預測方法為將驗證資料輸入至森林中每一棵迴歸樹中，每棵樹皆將得到一個預測數值，最終預測結果為所有樹預測數值之平均，表示式如下：

$$Predict = ave(\sum_{i=1}^{N_{trees}} predict_i) \quad (2)$$

3.2 支撐向量機

支撐向量機(Support Vector Machine)是由 Vapnik 與其共同研究者以統計學習理論為基礎所提出來的一個機器學方法(Vapnik, 1995;1998)。支撐向量機基於統計學習理論中結構風險最小化(structural risk minimization)的法則來處理多維度函數的分類與迴歸問題。

3.2.1 非線性支撐向量迴歸

假設資料序列$\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$，其中$x_i$迴歸函數的輸入向量，$y_i$為迴歸函數的輸出值，非線性向量迴歸係利用非線性映射函數(nonlinear mapping function) ϕ將非線性問題映射到高維度特徵空間(feature space)，使其變為線性問題，如圖 2 所示，因此迴歸函數為：

$$f(\mathrm{x}) = \boldsymbol{w} \cdot \phi_i + b \quad (3)$$

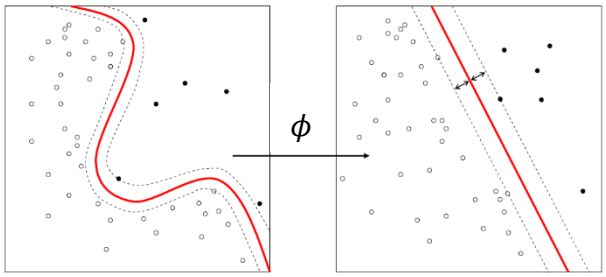

圖 2. 非線性映射函數(圖片來源：維基百科)

式中，**w**表示$f(\mathrm{x})$之平坦度(flatness)或複雜度(complexity)，**w**越小代表模式越不複雜，因此可以依據結構風險最小化法則，將線性支撐向量迴歸問題表示如下：

$$minimize \ \ \frac{1}{2}\boldsymbol{w}^T \cdot \boldsymbol{w} + C\sum_{i=1}^{l}(\xi_i - \xi_i^*) \quad (4)$$

$$\text{Subject to} \ \ y_i - (W^T \cdot x_i + b) \le \varepsilon + \xi_i$$
$$(W^T \cdot x_i + b) - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0, i = 1, \ldots, l$$

其中，$^1/_2\,\boldsymbol{w}^T \cdot \boldsymbol{w}$代表模式複雜度，$C\sum_{i=1}^{l}(\xi_i - \xi_i^*)$代表控制經驗風險，$\xi_i$與$\xi_i^*$為沉滯變數(slack variable)用以說明部分離群的訓練資料。

將所有沉滯變數的和乘上一個由使用者決定的成本參數(cost parameter)或稱為懲罰參數(penalty parameter) C，此參數越大代表誤差發生時對目標函數影響越大。另外，支撐向量迴歸主要基於 Vapnik's ε-insensitive loss function：

$$L_\varepsilon(y_i) = |y_i - f(x_i)|$$
$$= \begin{cases} 0 & \text{for } y_i - (W^T \cdot x_i + b) \le \varepsilon \\ |y_i - (W^T \cdot x_i + b)| - \varepsilon & \text{for } y_i - (W^T \cdot x_i + b) \ge \varepsilon \end{cases} \quad (5)$$

式中，$L_\varepsilon(y_i)$為支撐向量迴歸之損失函數，y_i為實際值，$f(x_i)$為估計值。由於定義一個可容許的誤差容忍區間(ε-tube)，若實際值落於此區間中，則損失函數(loss function)值為 0，否則損失函數則不為 0，此意義即為只有當落於誤差容忍區間外才給予懲罰，而經由損失函數即可定義出實際值與估計值之誤差，如圖一所示，由此也可得知只有落於容忍誤差範圍外的點才為支撐向量。

公式(3)中，主要藉由核技巧(kernel trick)，利用核函數 K 求解非線性問題，如下所示：

$$K(x_i, x_j) = \phi_i^T(x_i) \cdot \phi_i(x_j) \tag{6}$$

本研究使用幅狀基底函數(radial basis function)，其適用於多種不同資料特性之迴歸分析。

$$K(x_i, x_j) = exp(-\gamma|x_i - x_j|)^2 \tag{7}$$

上述γ為核函數之參數。最後獲得非線性支撐向量迴歸之決策函數為：

$$f(x) = \sum_{i,j=1}^{i}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b \tag{8}$$

3.3 參數最佳化

由上述方法簡介得知，隨機森林是由多顆(N_{trees})決策樹(每顆隨機選取m個變量)建立，故需兩個參數之設置，而支撐向量迴歸模式中則需要三個參數(C, γ, ε)訂定模式。為避免過度擬合之問題，本研究採用 10 摺交叉驗證(10 folds cross-validation)進行參數最佳化。以下簡單介紹參數最佳化之步驟。

3.3.1 隨機森林參數

本研究設定隨機森林決策數量$N_{trees} = 100$~2000與選取變量數量$m = 1$~19進行最佳參數測試，並配合使用 10 摺交叉驗證，將 834 筆資料隨機分成 10 等份，進行參數 10 摺的預報誤差分析，比較不同參數設定之誤差，以決定最佳參數。分析結果指出：森林大小約 500 顆時預報誤差漸趨穩定，而 m=6~7 時誤差變動不大，且 m=6 以上誤差值相差無幾。

3.3.2 支撐向量迴歸參數

本研究參考林智能(2003)網格搜尋法(grid-search method)，先以大網格(global)進行最佳參數區域搜尋($C = 2^0$~$2^{16}, \gamma = 2^0$~$2^{-14}, \varepsilon = 2^0$~$2^{-14}$)，再以小網格(local)搜尋最佳參數值，並配合交叉驗證，其作法如下：

1. 選定某組參數，例如：$(C, \gamma, \varepsilon) = (2^0, 2^0, 2^0)$，此組參數分別計算 10 摺資料中各摺預報誤差，計算平均誤差得此組參數之預報誤差。
2. 更換參數組數值重複上述分析($C = 2^0, 2^2 \ldots, 2^{14}, \gamma = 2^0, 2^{-2}, \ldots, 2^{-14}, \varepsilon = 2^0, 2^{-2}, \ldots, 2^{-14}$)，以選出誤差最小之最佳參數組合，例如：$(C, \gamma, \varepsilon) = (2^2, 2^{-2}, 2^{-4})$。
3. 最後，針對上述參數組合附近區間進行細網格(local)搜尋，例如：($C = 2^1, 2^{1.2} \ldots, 2^3, \gamma = 2^{-3}, 2^{-2.8}, \ldots, 2^{-1}, \varepsilon = 2^{-5}, 2^{-4.8}, \ldots, 2^{-3}$)，即可求得最佳參數組。

3.4 評鑑指標

本研究針對雨量預報效能之評鑑，本研究採取下列 4 項指標來評鑑之：

1. 均方根誤差(root mean squared error, RMSE)

$$RMSE = \sqrt{\frac{\sum_{t-1}^{n}(R_t - \hat{R}_t)^2}{n}} \tag{9}$$

其中，$\boldsymbol{R_t}$為觀測值(mm); $\widehat{\boldsymbol{R}}_{\boldsymbol{t}}$為預測值(mm)

2. 相關係數(correlation coefficient, CC)

$$CC = \frac{\sum_{t=1}^{n}(R_t - \bar{R}_t)(\hat{R}_t - \bar{\hat{R}}_t)}{\sqrt{\sum_{t-1}^{n}(R_t - \bar{R}_t)^2 \cdot \sum_{t-1}^{n}(\hat{R}_t - \bar{\hat{R}}_t)^2}} \tag{10}$$

其中，$\boldsymbol{R_t}$為觀測值(mm); $\widehat{\boldsymbol{R}}_{\boldsymbol{t}}$為預測值(mm)

3. 效率係數(coefficient of efficiency, CE)

$$CE = 1 - \frac{\sum_{t=1}^{n}[R_t - \hat{R}_t]^2}{\sum_{t=1}^{n}[R_t - \bar{R}_t]^2} \tag{11}$$

其中，$\boldsymbol{R_t}$為觀測值(mm); $\widehat{\boldsymbol{R}}_{\boldsymbol{t}}$為預測值(mm)

4. 最大絕對誤差(maximum absolute error, MAE)

$$MAE = \max\{|R_t - \hat{R}_t|\} \tag{12}$$

其中，$\boldsymbol{R_t}$為觀測值(mm); $\widehat{\boldsymbol{R}}_{\boldsymbol{t}}$為預測值(mm)

四、結果與討論

4.1 雨量預報模式

本研究考量雨量、颱風因子以及天氣因子，進行宜蘭河上游三站未來 1~3 小時之雨量預報，整個模式關係可由下列公式表示：

$$R(t+T) = f[R(t-1), R(t-2), TY_1(t-1), \dots, TY_n(t-1), L_1(t-1), \dots L_n(t-1)] \tag{13}$$

其中，R 為雨量；TY 為颱風特性因子；L 為集水區天氣因子。

4.1.1 雨量預報結果

為瞭解不同輸入因子於雨量預報之表現，本研究比較(1)僅考慮雨量因子(SVM Rain-only)與(2)考慮雨量因子、颱風因子以及天氣因子(SVM NEW)於雨量預報之表現，採用支撐向量迴歸進行預報，結果發現納入颱風因子與天氣因子後，長延時預報得到大幅改善，以 1996 賀伯颱風大礁溪站為例(圖 3)，僅考慮雨量因子於未來 2~3 小時之預報結果表現較不理想，而新因子加入後雖尚有預報延遲，但預報趨勢卻改善許多。

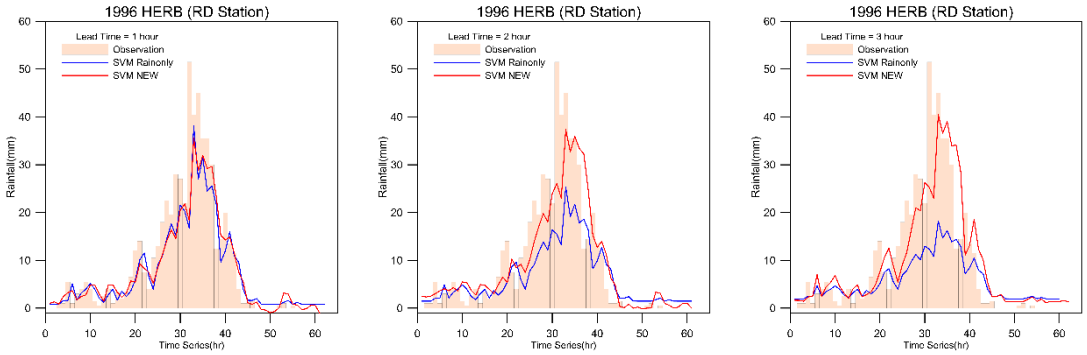

圖 3. 1996 賀伯颱風大磧溪站雨量預報(大磧溪站)

經由上述的比較後，本研究進一步比較隨機森林與支撐向量機於雨量預報的表現。以 1996 大磧溪站賀伯颱風事件模擬結果(圖 4)為例，相較觀測雨量，隨機森林之預報雨量無時間延遲之現象，於未來 1 小時預報雨量之峰值較接近觀測雨量峰值。

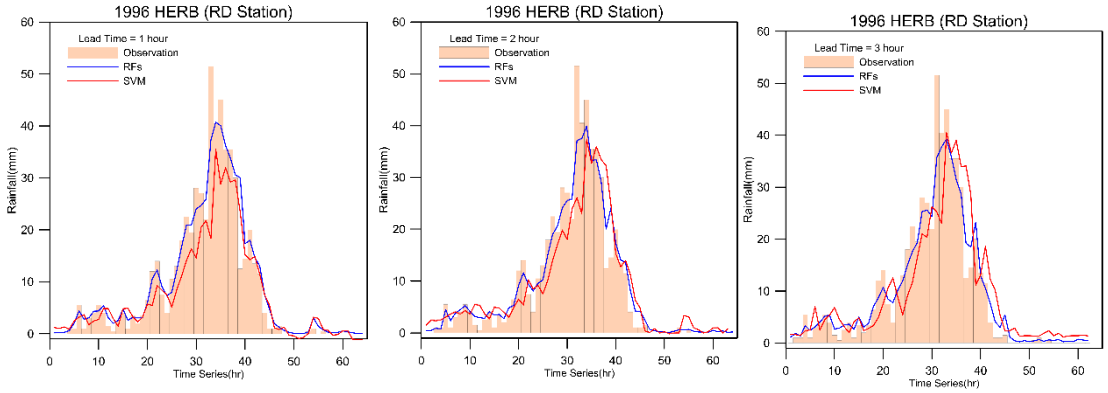

圖 4. 隨機森林與支撐向量機雨量預報比較圖(大磧溪站)

於評鑑指標的部分，以大磧溪站為例，隨機森林於未來 1 小時之雨量預報，其評鑑指標分別為：CC=0.94、RMSE=4.08、CE=0.76、MAE=31.05，相較於支撐向量機皆有明顯的改善。隨機森林於未來 2~3 小時之雨量預報亦有不錯之表現，反觀支撐向量機雨量預報之表現則會隨著預報時間加長而逐漸降低。整體而言，隨機森林於雨量預報較支撐向量機有更佳的表現。

表 1. 評鑑指標(大磧溪站)

RFs	CC	RMSE	CE	MAE
1HOUR	0.94	4.08	0.76	31.05
2HOUR	0.93	4.15	0.72	33.25
3HOUR	0.93	4.40	0.70	37.81
SVM	CC	RMSE	CE	MAE
1HOUR	0.75	7.37	-0.04	59.37
2HOUR	0.65	8.29	-0.34	71.10
3HOUR	0.55	9.48	-0.27	77.86

4.2 水位預報模式

本研究檢視雨量預報結果後，選用較好雨量預報模式隨機森林，建立水位預測模式。輸入因子為各雨量站 t-1~t-4 時刻之時雨量資料及水位站

t-1~t-2 時刻之時水位資料，預測 t 時刻之水位，如下式：

$$\hat{S}_{S,R}(t) = f[S(t-1), S(t-2), R_i(t-1), R_i(t-2), R_i(t-3), R_i(t-4)] \quad (14)$$

其中，$\hat{S}_{S,R}(t)$為 t 時刻下之預測水位；$S(t-1)$為 t-k 時刻下之觀測水位；$R_i(t-1)$為第 i 個雨量站 t-k 時刻下之觀測雨量；i 為雨量站站數，即 i=1~3。

4.2.1 水位預報結果

圖 5 為 2004 艾利颱風期間水位預報之結果，本研究以隨機森林進行未來 1~3 小時水位預測，並比較(1)僅考慮雨量因子(Old Simulation)之預報雨量於水位預報與(2)考慮雨量、颱風及天氣因子(New Simulation)之預報雨量於水位預報之表現，結果顯示：後者較無時間延遲之問題，預報亦較為精準。

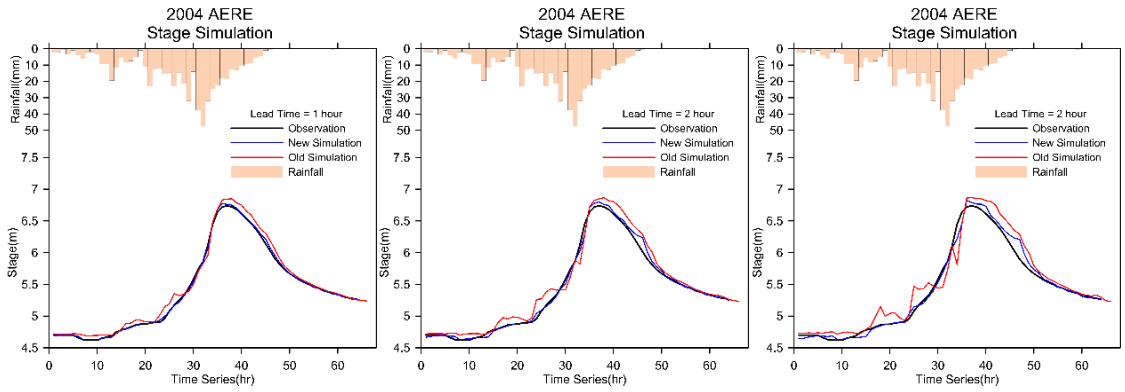

圖 5. 隨機森林水位模擬比較圖(中山水位站)

4.3 結論

1. 本研究為瞭解不同輸入因子於雨量預報之表現，以不同輸入因子建構雨量預報模式，結果顯示：僅考慮雨量因子之雨量預報具有延遲及不準確的問題；本研究引入颱風因子及天氣因子進行雨量預報，結果發現加入新因子的雨量預報有著顯著的改善。此外，本研究亦使用隨機森林進行雨量預報，其預報結果更為準確，亦成功解決雨量預報延遲的問題。
2. 本研究修正雨量預報後，將考慮雨量、颱風及天氣因子之預報雨量引入水位預報，以隨機森林建立模式，結果顯示：預報模式大幅增加長延時水位預報的準確性，並改善水位延遲預報問題。
3. 本研究建議未來可採用隨機森林之重要變量分析，嘗試將某些不敏感因子去除，以重要變量進行雨量或水位預報，加強預報模式表現。

參考文獻

陳憲宗，2006，支撐向量機及模糊推理模式應用於洪水水位之及時機率預報，國立成功大學博士論文

郭家妏，2014，隨機森林在河川水位即時預報之應用，國立成功大學碩士論文

Hsu, C.-W., Chang, C.-C. and Lin, C,-J., 2003, A Practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~cjlin

Verikas, A., Gelzinis, A. and Bacauskiene, M., 2011, Mining data with random forests: A survey and results of new tests, Pattern Recognition, 44, 330-349.