

NGHIÊN CỨU THỬ NGHIỆM DỰ BÁO XÁC SUẤT ĐỊNH LƯỢNG MƯA DỰA TRÊN PHƯƠNG PHÁP THỐNG KÊ SAU MÔ HÌNH TỪ SẢN PHẨM DỰ BÁO CỦA MÔ HÌNH HRM VÀ GSM

Phần I: Xây dựng các mô hình thống kê

TS. Lê Đức, NCS. Võ Văn Hòa, ThS. Đỗ Lệ Thủy, TS. Bùi Minh Tăng
ThS. Dư Đức Tiến, CN Nguyễn Mạnh Linh, CN. Nguyễn Thanh Tùng
Trung tâm Dự báo KTTV Trung ương

Bài báo này trình bày một số kết quả nghiên cứu ứng dụng các phương pháp thống kê bao gồm hồi quy tuyến tính đa biến (MLR), hồi quy logistic (LR), mạng thần kinh nhân tạo (ANN) và phân tích phân biệt Fisher (FDA) cho bài toán dự báo xác suất định lượng mưa (PQPF). Các mô hình thống kê này được xây dựng tách biệt cho từng tập nhân tố dự báo là các sản phẩm dự báo từ mô hình HRM và GSM, và áp dụng cho 130 trạm quan trắc khí tượng bề mặt dựa trên chuỗi số liệu từ năm 2003 đến năm 2009. Phần I của bài báo sẽ giới thiệu khái quát về phương pháp xây dựng các mô hình MOS nói trên từ khâu thiết kế bộ nhân tố sơ tuyển, kiểm tra chất lượng thăm sát, xử lý số liệu và tuyển chọn nhân tố và phương pháp đánh giá kết quả. Các kết quả nghiên cứu và đánh giá sẽ được trình bày trong phần II của bài báo.

1. Mở đầu

Dự báo mưa, đặc biệt dự báo định lượng mưa là một vấn đề rất khó khăn đồng thời cũng là một trong những yêu cầu cấp thiết trong công tác dự báo, đặc biệt trong dự báo bão, lũ, phục vụ phòng tránh thiên tai và kinh tế xã hội, điều tiết hồ chứa. Việc sử dụng phương pháp synop truyền thống chỉ có thể dự báo mưa một cách định tính. Tuy các mô hình dự báo số trị có thể đưa ra những dự báo định lượng, nhưng những sản phẩm dự báo trực tiếp từ mô hình (DMO - Direct Model Output) thường có những sai số hệ thống nhất định liên quan đến nhiều yếu tố, đặc biệt là vấn đề tham số hóa đối với sản phẩm mưa trực tiếp từ mô hình. Hơn nữa, DMO cũng chưa tính đến các yếu tố địa phương của chính những địa điểm dự báo. Từ lâu, phương pháp thống kê bằng cách xây dựng các phương trình hồi quy tuyến tính đơn

giản giữa yếu tố dự báo với các yếu tố đã biết (nhưng không phải từ mô hình số trị) đã được áp dụng trong dự báo thời tiết. Từ khi mô hình số được đưa vào nghiệp vụ, phương pháp thống kê có kết hợp với các sản phẩm của mô hình bắt đầu được hình thành để làm tăng cường thêm những kết quả của dự báo động lực trong dự báo nghiệp vụ. Hai phương pháp thống kê cơ bản có kết hợp sản phẩm của mô hình là PP (Perfect Prognosis) và MOS (Model Output Statistics). Cả hai phương pháp tiếp cận này đều sử dụng các phương trình hồi quy nhiều biến, có lọc. Điểm khác biệt giữa hai phương pháp này là cách sử dụng các tham số dự báo từ mô hình. PP chỉ sử dụng các tham số dự báo của mô hình khi thực hiện dự báo, trong khi MOS sử dụng các tham số này trong cả khi phát triển các phương trình thống kê và thực hiện dự báo.

Hiện nay, ở hầu hết các trung tâm dự báo trên thế giới, song song với việc đưa ra các sản phẩm mô hình thường có các chỉ dẫn thống kê để thêm giá trị vào những sản phẩm trực tiếp của mô hình. Đây là phương thức hợp lý để diễn xuất mô hình một cách khách quan nhằm loại bỏ những sai số của mô hình và dự báo định lượng cho địa điểm có tính đến các điều kiện khí hậu và địa phương cụ thể. Mỹ là nơi đã áp dụng MOS nhiều nhất đối với các mô hình khác nhau để đưa ra các sản phẩm dự báo nghiệp vụ như: NGM - MOS dự báo cho khoảng 600 địa điểm trên nước Mỹ cho thời hạn 6 đến 60 giờ; ETA - MOS dự báo cho 1250 địa điểm với thời hạn dự báo từ 6 đến 60 giờ. Ngoài ra, Trung tâm Dự báo Hạn vừa Châu Âu (ECMWF) cũng sử dụng phương pháp MOS đối với sản phẩm mô hình của mình. Australia cũng áp dụng phương pháp MOS đối với mô hình nghiệp vụ GAPS và LAPS cho hơn 600 địa điểm dự báo với thời hạn đến 7 ngày. Cả PP và MOS cũng được sử dụng trong nghiệp vụ ở Trung tâm Dự báo Bắc Kinh (Trung Quốc) từ năm 1984, dự báo cho khoảng 260 địa điểm với thời hạn 24 đến 60h. Tuy nhiên, việc sử dụng các dự báo bằng PP hay MOS vẫn còn đang tiếp tục được đánh giá và hoàn thiện hơn. Gần đây, cùng với sự phát triển của tốc độ tính toán và xử lý của máy tính, những phương pháp hiện đại hơn bắt đầu được quan tâm chú ý và thử nghiệm như SSM, ANN, FDA, UMOS, lọc Kalman, ...

Bài báo này trình bày một số kết quả nghiên cứu ứng dụng các phương pháp thống kê bao gồm hồi quy tuyến tính đa biến (MLR), hồi quy logistic (LR), mạng thần kinh nhân tạo (ANN) và phân tích phân biệt Fisher (FDA) cho bài toán dự báo xác suất định lượng mưa (PQPF). Phần I của bài báo này sẽ chỉ tập trung trình bày về phương pháp xây dựng các mô hình thống kê để dự báo PQPF từ tuyển chọn bộ nhân tố dự tuyển, kiểm tra chất lượng thám sát, xử lý số liệu và tuyển chọn nhân tố và phương pháp

đánh giá. Các kết quả nghiên cứu và đánh giá sẽ được trình bày trong phần II của bài báo.

2. Thiết kế và xây dựng các mô hình MOS dự báo PQPF

a. Thiết kế các mô hình MOS

Như đã biết, do lượng mưa thực tế thường không có được một tương quan tốt với các đại lượng dự báo từ mô hình, các nghiên cứu trên thế giới thường chỉ áp dụng MOS cho bài toán PQPF. Bởi vậy, nghiên cứu này sẽ tập trung chủ yếu vào PQPF với 4 phương pháp: hồi quy tuyến tính đa biến (MLR), hồi quy logistic (LR), phân tích phân biệt Fisher (FDA) và mạng thần kinh nhân tạo (ANN). Mô tả lý thuyết của các phương pháp này đã được trình bày trong nhiều tài liệu và sách tham khảo nên sẽ không được trình ở đây. Bốn phương pháp này sẽ được sử dụng để xây dựng các mô hình PQPF cho một số ngưỡng mưa cho trước với bộ nhân tố đầu vào riêng biệt từ mô hình HRM phiên bản 14km đang chạy nghiệp vụ tại TTDBTU' và mô hình GSM của Cơ quan Khí tượng Nhật Bản (JMA).

Các phương trình dự báo sẽ được xây dựng riêng biệt cho mỗi điểm trạm và không xây dựng một phương trình chung cho các điểm trạm trên một khu vực thuần nhất về mặt khí hậu bằng cách tổ hợp các trạm khi xây dựng phương trình thống kê. Theo Applequist và cộng sự (2002), với tập số liệu quan trắc đủ dài tại mỗi điểm trạm, không nên xây dựng một phương trình chung cho mọi điểm trạm trong một khu vực bởi chất lượng dự báo luôn kém hơn so với chất lượng dự báo có được nếu xây dựng phương trình thống kê cho mỗi trạm. Các mô hình PQPF thử nghiệm sẽ được xây dựng cho toàn bộ các điểm trạm đo mưa trên lãnh thổ Việt Nam nếu chuỗi số liệu đo mưa đủ dài để xây dựng phương trình thống kê. Hiện tại, TTDBTU' nhận được hàng ngày số liệu từ 170 trạm khí tượng bề mặt. Trong số này sẽ chỉ có 130 trạm khí tượng đạt được yêu cầu trên (xem bảng 1).

Bảng 1. Số lượng các trạm quan trắc khí tượng bề mặt của 9 khu vực hành chính địa lý được thử nghiệm dự báo theo MOS

TT	Tên khu vực hành chính địa lý	Số lượng các trạm quan trắc khí tượng bề mặt được sử dụng
1	Khu vực Đồng bằng Bắc Bộ	12 trạm khí tượng
2	Khu vực Đông Bắc	16 trạm khí tượng
3	Khu vực Tây Bắc	13 trạm khí tượng
4	Khu vực Việt Bắc	15 trạm khí tượng
5	Khu vực Bắc Trung Bộ	16 trạm khí tượng
6	Khu vực Trung Trung Bộ	15 trạm khí tượng
7	Khu vực Nam Trung Bộ	11 trạm khí tượng
8	Khu vực Tây nguyên	12 trạm khí tượng
9	Khu vực Nam Bộ	21 trạm khí tượng

Yếu tố dự báo với bài toán PQPF sẽ là biến nhị phân (có hoặc không) dựa trên quan trắc mưa thực tế với 6 ngưỡng mưa được lựa chọn 1, 2, 5, 10, 20, 50mm/24h và giá trị dự báo tương ứng là xác suất mưa thuộc khoảng $[0, 1]$. Việc lựa chọn các ngưỡng mưa này là dựa trên cơ sở khoa học của rất nhiều nghiên cứu về PQPF trên thế giới cũng như gợi ý của Tổ chức Khí tượng thế giới (WMO). Tại mỗi điểm trạm, phương trình thống kê sẽ được phát triển độc lập cho 2 mùa: mùa hè (từ 15/4 cho đến 14/10) và mùa đông (từ 15/10 năm trước đến 14/4 năm sau). Đây là cách phân chia được rất nhiều nghiên cứu sử dụng. Mỗi mùa có một tập dữ liệu 5 năm quan trắc và dự báo. Thực hiện kỹ thuật kiểm tra chéo, 4 năm số liệu sẽ được sử dụng như tập số liệu phụ thuộc để phát triển mô hình, năm số liệu còn lại được sử dụng như tập số liệu độc lập trong đánh giá. Kết quả đánh giá do đó sẽ có tính thuyết phục hơn, đảm bảo về tính thống kê, khi cả 5 năm dữ liệu đều được đưa vào đánh giá.

Có hai hướng tiếp cận với cách xác định nhân tố dự báo. Theo cách tiếp cận truyền thống kể từ khi Glahn và Lowry (1972) đề xuất phương pháp MOS, nhân tố dự báo bao gồm tất cả các biến dự báo thích hợp được xác định tại điểm trạm. Do đó, các trường dự báo trên lưới cần được nội suy về điểm trạm theo một phương pháp thích hợp. Khi nội suy

như vậy (ngoại trừ nội suy đơn giản bằng cách lấy điểm gần nhất), ta đã làm trơn trường dự báo dẫn đến mất mát một số thông tin. Đây chính là lý do một số tác giả như Jacks và cộng sự (1990) không thực hiện nội suy về trạm mà lấy biến dự báo từ 4 điểm lưới bao quanh điểm trạm làm nhân tố dự báo. Theo hướng tiếp cận như vậy, ta có thể sử dụng toàn bộ điểm lưới của trường dự báo làm nhân tố dự báo. Tuy nhiên, số nhân tố dự báo sẽ trở nên rất lớn dẫn đến những khó khăn khi thực hiện. Để giảm thiểu số nhân tố dự báo theo cách này, người ta chỉ lựa chọn một số biến quan trọng và thực hiện các phép biến đổi toán học trên các trường dự báo với mục đích giữ lại những thông tin quan trọng. Trong nghiên cứu này, chúng tôi chủ yếu sử dụng nhân tố dự báo theo cách truyền thống với giá trị dự báo tại các điểm trạm.

b. Lựa chọn và xây dựng bộ nhân tố sơ tuyến

Số liệu quan trắc được sử dụng xây dựng các phương trình thống kê bao gồm số liệu mưa tích lũy 24 giờ trên 130 trạm khí tượng trên lãnh thổ Việt Nam từ năm 2003 cho đến tháng 4 năm 2009. Quan trắc ngoài việc được sử dụng như yếu tố dự báo còn được sử dụng dưới dạng nhân tố dự báo. Nhân tố dự báo này được cho dưới dạng nhị phân cho biết 24 giờ trước đó lượng mưa thực tế có xảy ra hay không theo một số ngưỡng mưa cho trước 01, 02,

05, 10, 20, 50mm. Ví dụ với lượng mưa 9mm, giá trị của nhân tố dự báo ứng với ngưỡng mưa trên lần lượt là 1, 1, 1, 0, 0, 0. Đưa thêm các giá trị này vào quá trình xây dựng phương trình thống kê nhằm đảm bảo tính thống nhất hoặc quán tính (persistence) cũng được sử dụng trong phương trình thống kê.

Số liệu phân tích và dự báo từ 2 mô hình GSM và HRM tại các thời điểm 00, +12h, +24h sẽ được sử dụng với thời gian tương ứng với thời gian của số liệu quan trắc mưa. Số liệu này được cho trên lưới kinh vĩ độ phân giải 1.25° với GSM và 0.125° với HRM. Dựa theo đề xuất của Glahn và Lowry (1972), Applequist và cộng sự (2002), Wilson và Vallee (2002), các biến có thể trở thành nhân tố dự báo cần được rút ra từ dự báo mô hình bao gồm: mưa, mưa đối lưu, tốc độ mưa, tốc độ mưa đối lưu, lượng mây, gió 10m, nhiệt độ 2m, độ hụt điểm sương 2m, áp suất mực biển, độ cao chân mây, độ cao đỉnh mây, địa thế vị, gió, nhiệt độ, độ ẩm, tốc độ thẳng đứng tại các mực áp suất cơ bản 1000, 950/925 (HRM/GSM), 850, 700, 500, 300mb. Từ các biến cơ sở này, các tác giả trên đề xuất một số biến dẫn xuất cũng có khả năng trở thành nhân tố dự báo. Dưới đây là các biến dẫn xuất nhóm nghiên cứu sử dụng:

* Trường trên cao: độ ẩm riêng, độ ẩm tương đối, nhiệt độ điểm sương, nhiệt độ thế vị tương đương, xoáy tương đối, phân kỳ, hội tụ ẩm.

* Gradient: áp dụng toán tử gradient theo 2 hướng kinh độ, vĩ độ ta sẽ có các trường gradient cho địa thế vị, gió, nhiệt độ, độ hụt điểm sương. Riêng với phương thẳng đứng, toán tử gradient cho ta tỉ lệ giảm theo chiều cao với nhiệt độ, độ đứt với hai thành phần gió. Với các biến còn lại gồm: địa thế vị, nhiệt độ thế vị tương đương, độ ẩm riêng: ta chỉ xác định độ chênh lệch giữa các biến tại các mực áp suất khác nhau như độ lệch của phân kỳ giữa mực 850 và 300mb.

* Bình lưu: áp dụng toán tử bình lưu cho các biến nhiệt độ, độ ẩm riêng, độ hụt điểm sương, nhiệt độ thế vị tương đương, xoáy tương đối.

* Trung bình trong 1 cột khí quyển: tích phân thẳng đứng các trường 3 chiều từ mực 950mb đến

mực 500mb rồi lấy trung bình. Toán tử này được thực hiện cho tất cả các biến 3 chiều cơ sở cũng như dẫn xuất, cộng thêm các biến bình lưu.

* Các trường khác: các chỉ số bất ổn định (K và Showalter), lượng nước giáng thủy khả năng (xác định từ tích phân thẳng đứng của độ ẩm riêng), tích giữa tốc độ thẳng đứng và lượng nước giáng thủy khả năng, laplacian của bình lưu nhiệt độ.

Có một số biến không tồn tại trong kết quả dự báo của GSM (do JMA không phát báo các biến này): mưa đối lưu, tốc độ mưa, tốc độ mưa đối lưu, lượng mây, cao chân mây, độ cao đỉnh mây và tốc độ thẳng đứng. Do đó, số lượng các biến có thể trở thành nhân tố dự báo với GSM sẽ nhỏ hơn so với số biến có thể với HRM.

c. Phương pháp xử lý số liệu và tuyển chọn nhân tố

Để đảm bảo chất lượng số liệu quan trắc trước khi đưa vào sử dụng, số liệu mưa được so sánh với số liệu mưa từ vệ tinh TRMM thông qua 4 điểm lưới TRMM xung quanh điểm trạm. Nếu lượng mưa tại trạm chênh lệch quá lớn so với lượng mưa TRMM từ 4 điểm này, quan trắc sẽ bị loại bỏ. Ngoài ra, nếu lượng mưa vượt quá giá trị khí hậu 5 σ (do mưa thường có phân bố Gamma thay vì phân bố Gauss), quan trắc tại trạm sẽ được kiểm tra với các thông tin lịch sử, đảm bảo mưa lớn thực sự diễn ra hay không.

Do phương pháp LR khi xác lập phương trình thống kê luôn cho giá trị xác suất mưa dự báo trong khoảng [0,1], 2 phương pháp MLR và ANN có thể dự báo xác suất mưa nhỏ hơn 0 hay lớn hơn 1. Trong trường hợp này ta sẽ dùng kỹ thuật cắt (clipping) gán những giá trị dự báo nhỏ hơn 0 bằng 0 và lớn hơn 1 bằng 1. Riêng với phương pháp FDA, kỹ thuật xử lý phức tạp hơn do thực tế FDA chỉ phân lớp yếu tố dự báo thành 2 lớp có mưa hay không mưa theo ngưỡng cho trước mà không đưa ra dự báo xác suất giữa 0 và 1. Để thu được giá trị xác suất dự báo ta sẽ sử dụng kỹ thuật biến đổi (transforming) của Murphy và Winkler (1987). Giá trị phân lớp 0 và 1 từ FDA sẽ được thay thế bằng biến liên tục là khoảng cách từ điểm dự báo đến mặt phẳng

phân lớp. Tập các giá trị này, xác định từ FDA trên tập số liệu phụ thuộc, được sắp xếp theo thứ tự tăng dần và chia đều thành mười phần. Với mỗi phần, các quan trắc khi dự báo rơi vào phần này được tập hợp lại và xác định xác suất lượng mưa thực tế lớn hơn ngưỡng mưa đang xét tương ứng. Giá trị này được xem như xác suất dự báo sau đó khi áp dụng FDA trên tập dữ liệu độc lập và khoảng cách dự báo rơi vào phần mười tương ứng. Ví dụ khi khoảng cách dự báo trong khoảng [1,2] trên tập số liệu phụ thuộc, ta quan sát thấy tần suất mưa thực tế là 0.9. Giá trị 0.9 này sẽ là giá trị xác suất dự báo nếu khoảng cách dự báo trên tập dữ liệu độc lập sau đó thuộc khoảng [1,2].

Mặt khác, do nghiên cứu chủ yếu sử dụng bộ nhân tố dự báo là các giá trị dự báo tại các điểm trạm, nên các trường trên lưới cần được nội suy về điểm trạm theo phương pháp nội suy thích hợp. Với các trường trên cao, phương pháp nội suy tuyến tính được áp dụng. Với các trường bề mặt, giá trị dự báo được xác định đơn giản bằng điểm lưới gần nhất. Tuy nhiên, do trường dự báo từ GSM có độ phân giải khá thô 1.25° nên ngoại trừ trường mưa các biến bề mặt cũng được nội suy tuyến tính về điểm trạm. Như vậy, với 3 thời điểm 00, 12 và 24, mỗi điểm trạm sẽ có 726 biến dự báo với HRM và 666 với GSM. Kết hợp với tập các nhân tố dự báo lấy từ lượng mưa quan trắc trong quá khứ tại thời điểm dự báo đã nói đến ở trên, sẽ có khoảng 700 nhân tố dự báo với mỗi mô hình.

Một vấn đề khác liên quan đến bộ nhân tố dự báo ban đầu cần phải giải quyết là sự khác nhau về thứ nguyên và bậc đại lượng. Để có một tập nhân tố dự báo với bậc độ lớn tương đương và có thứ nguyên như nhau, các nhân tố dự báo cần được chuẩn hóa trước khi xây dựng phương trình thống kê. Điều này khá hiển nhiên, nếu quan sát các biến dự báo ở trên có thể thấy với một tập lớn các biến dự báo được đưa vào tuyển chọn, bậc độ lớn biến thiên từ 10-12 với laplacian của bình lưu nhiệt độ tới 104 với độ cao địa thế vị tại mực 300mb. Tất nhiên các biến này khác nhau về thứ nguyên và không thể so sánh với nhau, nhưng phương trình thống kê xem các nhân tố dự báo như nhau, không phân biệt

thứ nguyên. Do đó, chênh lệch quá lớn về độ lớn giữa các nhân tố dự báo sẽ dẫn đến sai sót trong xây dựng phương trình thống kê khi các đại lượng quá lớn hay quá nhỏ sẽ bị bỏ qua. Để khắc phục vấn đề này, tập hợp các nhân tố dự báo sẽ được chuẩn hóa về một tập hợp nhân tố mới theo công thức như sau:

$$\hat{x}_k = \frac{x_k - \bar{x}_k}{sd_k} \quad (1)$$

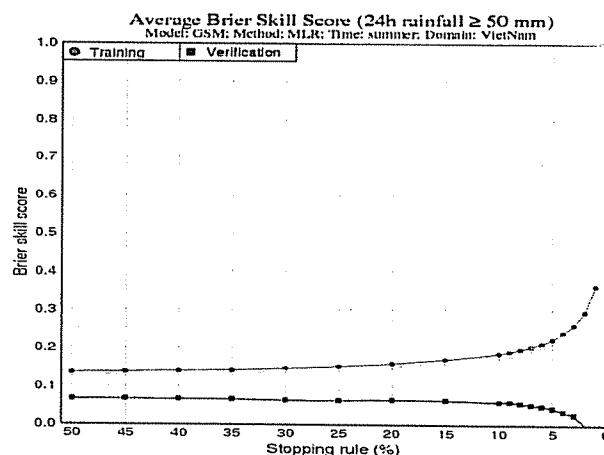
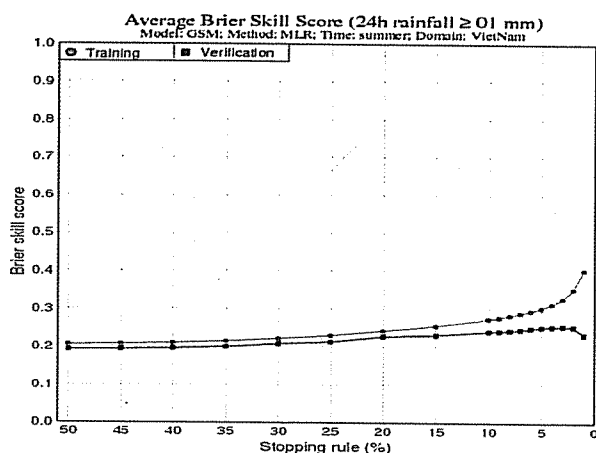
Ở đây, $\hat{x}_k$ là giá trị chuẩn hóa của nhân tố dự báo thứ x_k , $\bar{x}_k$ và sd_k tương ứng là giá trị trung bình và độ lệch chuẩn của nhân tố x_k xác định từ tập số liệu phụ thuộc. Sau khi được chuẩn hóa, các nhân tố mới $\hat{x}_k$ là vô thứ nguyên và phần lớn có giá trị tập trung trong khoảng từ -3 cho đến +3. Đối với các biến luôn dương (như lượng mưa, độ ẩm, ... thì các giá trị này sẽ nằm trong khoảng [0,3].

Với một tập khoảng 700 nhân tố dự báo cho mỗi điểm trạm, khối lượng tính toán sẽ rất lớn và cũng không cần thiết khi các biến có quan hệ tương hỗ nhất định. Do đó, một quá trình tuyển chọn nhân tố cần được thực hiện nhằm giảm thiểu khối lượng tính toán trong khi vẫn đảm bảo những thông tin quan trọng thu được từ các biến dự báo. Với bài toán PQPF, nhóm nghiên cứu sẽ thực hiện tuyển chọn theo phương pháp tuyển chọn từng bước tiến với nguyên lý tương đương như phương pháp hồi quy từng bước tiến. Tuy nhiên, thay vì hệ số tương quan như phương pháp hồi quy từng bước, chỉ tiêu tuyển chọn được sử dụng là chỉ số kỹ năng Brier (BSS).

Tại mỗi bước tuyển chọn, lần lượt các nhân tố sẽ được đưa vào xây dựng phương trình thống kê (với một phương pháp thống kê nào đó) trên tập số liệu phụ thuộc. Nhân tố tương ứng với phương trình thống kê có BSS lớn nhất sẽ được giữ lại và đưa vào danh sách tuyển chọn. Tại bước tiếp theo, quá trình được lặp lại nhưng chỉ thực hiện tuyển chọn trên các nhân tố chưa được tuyển chọn. Mỗi nhân tố mới này sẽ được kết hợp với các nhân tố đã tuyển chọn trước đó để xây dựng phương trình thống kê mới. Quá trình sẽ dừng lại khi BSS không tăng thêm với một tỷ lệ phần trăm cho trước so với BSS của bước trước đó. Tỷ lệ này được gọi là tỷ lệ dừng

(stopping rule), đóng vai trò quan trọng quyết định tới chất lượng dự báo. Nếu tỷ lệ dừng quá nhỏ, số lượng các nhân tố dự báo được chọn sẽ lớn và hiện tượng quá khớp (overfitting) có thể xảy ra. Mặc dù BSS xác định từ tập số liệu phụ thuộc có giá trị lớn nhưng BSS xác định từ tập số liệu độc lập sẽ có giá trị nhỏ, thậm chí âm. Nếu tỷ lệ dừng quá lớn, một số thông tin dự báo quan trọng sẽ không được đưa vào phương trình thống kê và BSS sẽ có giá trị nhỏ. Để xác định tỷ lệ dừng tối ưu, ta sẽ hiển thị đồng thời giá trị BSS tính từ tập số liệu phụ thuộc và tập số liệu độc lập dựa theo các tỷ lệ dừng khác nhau như trên hình 1.

Từ hình 1 có thể thấy BSS xác định trên tập số liệu phụ thuộc tăng dần khi tỷ lệ dừng giảm dần nhưng BSS xác định trên tập số liệu độc lập bắt đầu giảm (còn âm khi tỷ lệ dừng vượt quá 2% ở ngưỡng 50mm) khi tỷ lệ dừng vượt quá 3% với ngưỡng 01mm và 15% với ngưỡng 50m. Tỷ lệ dừng 3% và 15% do đó có thể xem là tỷ lệ dừng tối ưu trong những trường hợp này. Như vậy không có một tỷ lệ dừng duy nhất tối ưu cho mọi ngưỡng. Tỷ lệ dừng sẽ tăng dần khi ngưỡng mưa tăng dần và các tỷ lệ này cũng khác nhau cho các mô hình, phương pháp và thời gian dự báo khác nhau.



Hình 2. Sự biến thiên của chỉ số BSS trung bình trên khu vực Việt Nam theo các giá trị tỷ lệ dừng khác nhau với phương pháp MLR từ mô hình GSM dự báo cho mùa hè với 2 ngưỡng mưa 01mm (trái) và 50mm (phải)

Phương pháp tuyển chọn từng bước như trên được áp dụng với cả 3 phương pháp thống kê MLR, LR và FDA trong đó 2 phương pháp LR và FDA đòi hỏi một khối lượng tính toán lớn hơn rất nhiều so với phương pháp MLR. Riêng với phương pháp ANN do khối lượng tính toán rất lớn liên quan đến quá trình cực tiểu hóa xác định các hệ số thống kê, khi tuyển chọn nhân tố trên một tập 700 nhân tố khả năng, khối lượng tính toán tăng lên rất đáng kể. Bởi vậy, nhóm nghiên cứu sẽ không đưa toàn bộ 700 nhân tố vào tuyển chọn mà sẽ rút gọn tập nhân tố khả năng trước khi tuyển chọn. Tập nhân tố này dựa trên tập nhân tố tuyển chọn theo phương pháp MLR với tỷ lệ dừng 1% do khối lượng tính toán theo phương pháp MLR không lớn.

d. Phương pháp đánh giá

Việc đánh giá dự báo PQPF từ các mô hình

thống kê thử nghiệm sẽ được thực hiện tại các điểm trạm dựa trên chỉ số kỹ năng Brier - BSS (Brier, 1950), biểu đồ tin cậy và đường đặc trưng hoạt động ROC (Wilks, 1995, 2006). Chi tiết về các chỉ số và biểu đồ đánh giá này có thể tham khảo trong nghiên cứu của Bùi Minh Tăng và cộng sự (2009). Tuy nhiên, nghiên cứu này sẽ không tập trung vào đánh giá tại từng điểm trạm mà phân vùng đánh giá dựa theo phân vùng hành chính địa lý (xem bảng 2.1). Đánh giá dựa theo phân vùng đảm bảo một tập dữ liệu đánh giá đủ lớn, không bị ảnh hưởng nhiều khi giá trị đo một trạm nào đó trong khu vực có sai số lớn. Hơn nữa, ta vẫn có thể quan sát phân bố của chất lượng dự báo theo không gian từ đó có những điều chỉnh về mặt mô hình hay khuyến cáo về chất lượng dự báo tại một khu vực cụ thể. Riêng với sự kiện hiếm như mưa lớn với ngưỡng 50mm, đánh giá

cần được thực hiện trên toàn khu vực Việt Nam.

Cuối cùng, để so sánh chất lượng dự báo từ 4 phương pháp theo 2 mô hình nhằm lựa chọn phương án tốt nhất thông qua chỉ số đánh giá BSS, ta cần kiểm nghiệm thực sự chất lượng dự báo từ phương pháp này có tốt hơn so với phương pháp kia hay không. Kiểm nghiệm thống kê cặp t sẽ được thực hiện (Applequist và cộng sự, 2002) dựa trên tập các độ lệch BSS tại cùng điểm trạm theo 2 phương pháp:

$$t = \frac{\bar{d}}{s/\sqrt{m}} \quad (2)$$

với $\bar{d}$ là độ lệch trung bình, s độ lệch chuẩn và m số điểm trạm. Khi m lớn hơn 30, t sẽ có phân bố chuẩn và ta dễ dàng kiểm tra độ tin cậy của giả thiết.

3. Kết luận

Bài báo này đã tiến hành xây dựng các mô hình thống kê dựa trên các kỹ thuật hồi quy tuyến tính đa biến (MLR), hồi quy logistic (LR), mạng thần kinh nhân tạo (ANN) và phân tích phân biệt Fisher (FDA) để dự báo xác suất định lượng mưa (PQPF). Các mô hình PQPF được xây dựng cho một số ngưỡng mưa cho trước với bộ nhân tố đầu vào riêng biệt từ mô hình HRM phiên bản 14km đang chạy nghiệp vụ tại TTDBTU và mô hình GSM. Các phương trình dự

báo sẽ được xây dựng riêng biệt cho mỗi điểm trạm và không xây dựng một phương trình chung cho các điểm trạm trên một khu vực thuần nhất về mặt khí hậu bằng cách tổ hợp các trạm khi xây dựng phương trình thống kê. Yếu tố dự báo với bài toán PQPF sẽ là biến nhị phân (có hoặc không) dựa trên quan trắc mưa thực tế với 6 ngưỡng mưa được lựa chọn 1, 2, 5, 10, 20, 50mm/24h và giá trị dự báo tương ứng là xác suất mưa thuộc khoảng [0,1].

Bộ nhân tố sơ tuyến sẽ bao gồm số liệu quan trắc đã được đưa về dạng nhị phân và các đại lượng dự báo từ mô hình HRM và GSM. Như vậy, với 3 thời điểm 00, 12 và 24, mỗi điểm trạm sẽ có 726 biến dự báo với HRM và 666 với GSM. Kết hợp với tập các nhân tố dự báo lấy từ lượng mưa quan trắc trong quá khứ tại thời điểm dự báo đã nói đến ở trên, sẽ có khoảng 700 nhân tố dự báo với mỗi mô hình và các nhân tố dự báo đều được chuẩn hóa trước khi xây dựng phương trình thống kê.

Để đánh giá kết quả dự báo PQPF từ các mô hình thống kê một cách khoa học và khách quan, các phương pháp đánh giá hiện đại và phù hợp sẽ được sử dụng bao gồm chỉ số kỹ năng Brier, biểu đồ tin cậy và đường đặc trưng hoạt động ROC. Các kết quả nghiên cứu và đánh giá sẽ được trình bày chi tiết trong phần II của bài báo.

Tài liệu tham khảo

1. Bùi Minh Tăng và cộng sự, 2009: Nghiên cứu thử nghiệm dự báo định lượng mưa từ sản phẩm dự báo của mô hình HRM và GSM. Báo cáo tổng kết Đề tài NCKH cấp Bộ, 127 trang.
2. Applequist, S., Gahrs G. E., Pfeffer R. L. and Niu X. F., 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting*, 17, 783-799.
3. Glahn H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. App. Meteor.*, 11, 1203-1211.
4. Jacks E., J. B. Bower, V. J. Dagostaro, J. P. Dallvalle, M. C. Erickson and J. C. Su, 1990: New NGM based MOS guidance for Maximum/Minimum Temperature. Probability of Precipitation, Clud Amount, and Surface Wind. *Wea. Forecasting*, 5, 128-138
5. Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.
6. Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, Second Edition, 649 pp.
7. Wilson L. J., and M. Vallée, 2002: The Canadian Updatable Model Output Statistics (UMOS) system: Design and development test. *Wea. Forecasting*, 17, 206-222.