文章编号：1004-4965（2007）05-0491-06

# 基于小波分解和最小二乘支持向量机的西太平洋副高预测模型

刘科峰[1,2]，张韧[1,2]，于鹏[1]，王彦磊[1]，余丹丹[1]

（1. 解放军理工大学气象学院海洋环境与空间系，江苏 南京 211101；
2. 中国气象局广州热带海洋气象研究所，广东 广州 510080）

摘　　要：用小波分解（WT）和最小二乘支持向量机（LS-SVM）相结合的方法，建立西太平洋副热带高压面积指数的预报模型。该方法首先将西太平洋副热带面积指数（SI）分解为相对简单的带通分量信号，利用 LS-SVM 建立各分量信号的独立预报模型，然后对预报结果进行集成。为了评估和比较该方法的预报效果和技术优势，最后比较了在同等条件下 WT～LS-SVM 模型和神经网络、线性回归模型的独立检验预报效果。试验结果表明，该方法具有泛化能力强、预报精度高、训练速度快、稳定性好、便于建模等优点，具有良好的应用前景。

关　键　词：小波分解；最小二乘支持向量机；副热带高压
中图分类号：P456.7　　文献标识码：A

## 1 引　　言

西太平洋副热带高压（简称副高）是一个重要且复杂的天气系统，它是影响我国夏季天气气候的主要天气系统。我国夏季雨带的分布与移动和副高的季节性移动密切相关，它的强度和东西进退活动异常往往导致江淮流域的持续性的洪涝或干旱。对它的预报一直是人们非常关心但尚未取得满意结果的课题。

目前，基于自适应和非线性的人工神经网络方法在副高的研究和预测中取得一定的成功[1−2]。但是神经网络方法存在着难以克服的缺陷，如隐层单元神经元的数目难以确定；容易陷入局部最优；神经网络的结构设计依赖于设计者的先验知识和经验，缺乏一种有理论依据的严格设计程序等。另外，从概率统计的角度说，神经网络的学习算法采用经验风险最小化原理（ERM），仅仅试图使经验风险最小化，并没有使期望风险最小化，与传统的最小二乘法相比，在原理上缺乏实质性的突破，同时也缺乏理论依据。总之，神经网络学习算法缺乏定量的分析与机理完备的理论结果[3−4]。此外，副高的预测多见于月平均等大时间尺度的活动，基于中小时间尺度的副高预测活动的研究相对较少。因此有必要引入新理论、新方法对中小尺度的副高预测活动做更深入的研究。

1995 年，由贝尔实验室的 Vapnik 等人在统计学习理论的基础上提出了模式识别的新方法——支持向量机(Support Vector Machine，SVM)，它根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中，使结构风险最小，即同时最小化经验风险与 VC 维(Vapnik-Chervonenkis Dimensino)的界，以期获得最好的泛化能力。与经典支持向量机相比，最小二乘支持向量机(Least Square Support Vector Machine，LS-SVM)用等式约束代替不等式约束，求解过程变成了解一组等式方程，避免了求解耗时的二次规划(Quaratic Programing，QP)问题。对样本较少等某些问题求解较方便。相对于常用的不敏感损失函数[5]，LS-SVM 不再需要指定逼近的精度。

本文拟采用小波分析和最小二乘支持向量机相

收稿日期：2006-04-16；修订日期：2006-08-26
基金项目：国家自然科学基金项目（40375019）；热带海洋气象科学研究基金；热带季风重点开放实验室共同资助
作者简介：刘科峰，男，陕西人，在读博士研究生，研究方向：海洋气象。E-mail：fengke_liu@126.com

结合的方法，研究和探讨利用观测资料预测西太平洋副热带高压面积指数的途径方法。

# 2 统计学习理论和最小二乘支持向量机

## 2.1 统计学习理论

统计学习理论[5]提出机器学习的实际风险由两部分组成：一是经验风险 $R_{\text{emp}}(w)$，另一部分为置信范围，它与 VC 维及训练样本数有关。经验风险 $R_{\text{emp}}(w)$ 和实际风险 $R(w)$ 之间的关系可表示为

$$R(w) \le R_{\text{emp}}(w) + \phi(h/n) \tag{1}$$

式中 $h$ 为函数集的 VC 维；$n$ 为样本数。

式（1）表明在有限的训练样本下，学习机器的 VC 维越高则置信范围越大，导致真实风险与经验风险之间可能的差别越大。机器学习过程不但要使经验风险最小，还要使 VC 维尽量小以缩小置信范围，才能取得较小实际风险，即对未来的样本有较好的推广性。统计学习理论提出把函数集构造成一个函数子集序列，使各子集按照 VC 维的大小排列；在每个子集中寻找最小经验风险，在子集间折中考虑经验风险和置信范围，取得实际风险最小（图 1），这种思想称作结构风险最小化（Structural Risk Minimization，SRM），支持向量机就是 SRM 准则的具体体现。

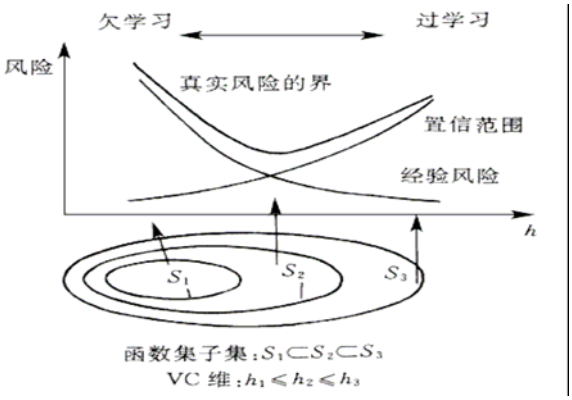


图 1 结构风险最小化示意图

## 2.2 最小二乘支持向量机

支持向量机方法是近年国际上开始流行的一种新颖的处理非线性分类和回归的有效方法。它以 Vapnik 提出的统计学习理论为基础，将样本空间映射到一个更高维以至于无穷维的特征空间，在特征空间中把寻求最优回归超平面问题归结为一个凸约束条件下的二次凸规划问题，从而求得全局最优解。Suykens 等[6]提出的最小二乘支持向量机是支持向量机的一种，它是将标准支持向量机算法中的不等式约束转化为等式约束而得到的。对非线性回归问题，设训练样本为 $(x_1, y_1), \cdots\cdots, (x_l, y_l) \in R^n \times R$。非线性回归函数为

$$f(x) = w^T \Phi(x) + b \tag{2}$$

对于最小二乘支持向量机，优化问题变为

$$\text{Min}\ \frac{1}{2}\|w\|^2 + \gamma \frac{1}{2}\left(\sum_i^p \xi_i^2\right) \tag{3}$$
$$s.t \quad y_i = w^T \Phi(x_i) + b + \xi_i \quad (\xi_i \ge 0,\ i = 1, \cdots\cdots, l)$$

求解式（3）的优化问题，可以引入 Lagrange 函数

$$L = \frac{1}{2}\|w\|^2 + \gamma \frac{1}{2}\left(\sum_{i=1}^p \xi_i^2\right) - \sum_{i=1}^l a_i (w^T \Phi(x_i) + b + \xi_i - y_i) \tag{4}$$

式中 $a_i$ 为 Lagrange 乘子；常数 $\gamma > 0$，它控制对超出误差的样本的惩罚程度。最优的 $a_i$ 和 $b$ 可以根据 KKT（Karush-Kuhn-Tuchker）条件得到

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \to w = \sum_{i=1}^l a_i \Phi(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \to \sum_{i=1}^l a_i = 0 \\ \dfrac{\partial L}{\partial \xi} = 0 \to a_i = \gamma \xi_i \\ \dfrac{\partial L}{\partial \xi} = 0 \to a_i = \gamma \xi_i \\ \dfrac{\partial L}{\partial a} = 0 \to w^T \Phi(x_i) + b + \xi_i - y_i = 0 \\ i = 1, \cdots\cdots, l \end{cases} \tag{5}$$

优化问题由式（5）转化为求解如下的线性方程

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & K(X_1, X_1) + 1/\gamma & K(X_1, X_2) & \cdots & K(X_1, X_l) \\ 1 & K(X_1, X_2) & K(X_2, X_2) + 1/\gamma & \cdots & K(X_2, X_l) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & K(X_l, X_1) & K(X_l, X_2) & \cdots & K(X_l, X_l) + 1/\gamma \end{bmatrix} \begin{bmatrix} b \\ a_1 \\ a_2 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix} \tag{6}$$

其中 $K(X_i, X_j)$ 为核函数。从而得到非线性回归函数的解为

$$f(x) = w^T \Phi(x) + b = \sum_{i=1}^{l} a_i K(X_i, X) + b \qquad (7)$$

最常用的核函数有多项式核函数、RBF 核函数、Sigmod 核函数等。

# 3　小波分解

1988 年 S Mallat 在构造正交小波基时提出了多分辨率分析（Multi～Resolution Analysis）概念，从空间的概念上形象地说明了小波的多分辨率特性，给出了正交小波的构造方法以及正交小波变换的快速算法，即 Mallat 算法[7]。小波分解旨在构造一个频率上高度逼近原始信号的正交小波基，这些频率分辨率不同的正交小波基相当于带宽各异的带通滤波器[8]。小波变换的多分辨率分析主要是对信号的低频空间作细致的分解，使其低频部分分辨水平越来越高，从而降低信号的复杂程度。

# 4　副高面积指数的 WT～LS-SVM 模型

## 4.1　资料说明

分析资料为美国国家中心（National Center for Environment Prediction，NCEP）和美国国家大气研究中心（National Center for Atmospheric Research，NCAR）提供的 1995 年 1 月～2004 年 12 月共 10 年的 2.5 °×2.5 °网格 500 hPa 逐日再分析高度场资料。统计 10 °N 以北，110～180 °E 范围内平均位势高度大于 586 dgam 网格点数，即可得 10 年逐候的副高面积指数（SI）序列（在天气分析中，一般取 500 hPa 上的 588 dgam 等高线来分析，但实际上有些年份画不出 588 线，故分析 586 线）。

## 4.2　WT～LS-SVM 模型的设计

10 年逐候的副高面积指数（SI（$t$），$t$=1，2，……，$n$）可看作一个复杂的信号，利用小波分解能够将复杂信号进行频率(周期)分离的特性，将副高面积指数序列{SI（$t$）}分解为相对简单的低频信号和高频信号，然后用 LS-SVM 分别对低频信号和高频信号建立预测模型，最后将预测得到的各频段信号进行重构，即可得到副高面积指数的最终预测值。

对高频信号和低频信号均选用 RBF 核函数 $K(x_i, x_j) = \exp(-\sigma \left\| x_i - x_j \right\|^2), \sigma > 0$ 构建 LS-SVM 模型。核函数确定后，还需确定两个相关的参数：$\sigma$、$\gamma$。其中 $\sigma$ 为核参数，调节核函数的平滑程度；$\gamma$ 为正则化参数，控制模型的复杂度和函数逼近误差的大小。这两个模型参数在很大程度上决定了该模型的学习能力及泛化能力。如何确定模型参数，目前尚缺乏一个客观有效的方法。我们采用逐步筛选的方法确定这两个模型参数：首先设置较大的参数取值范围，对参数进行大间隔步长的循环取值，通过训练和测试，根据预报结果与实际值的相关系数、平均绝对误差和相对误差的大小综合确定最优参数值，再以此参数值为中心，设置较小的参数范围，以小间隔步长重复上述步骤，直至最终确定出用于建立 LS-SVM 预报模型参数值，进而确定预报模型。

## 4.3　集成预测步骤

采用 WT～LS-SVM 对副高面积指数进行预测的步骤如下。

(1) 为防止数据溢出，同时加快运算速度，首先采用下式对副热带高压面积指数作归一化处理。

$$\overline{\mathrm{SI}}(t) = \frac{\mathrm{SI}(t) - \mathrm{SI}_{\min}}{\mathrm{SI}_{\max} - \mathrm{SI}_{\min}} \qquad (t=1，\cdots\cdots，6，\cdots\cdots，n)$$

式中 SI($t$) 为副高面积指数值，$\overline{\mathrm{SI}}(t)$ 为归一化后的副高面积指数值，$\mathrm{SI}_{\max}$、$SI_{\min}$ 为副高面积指数序列的最大、最小值。

(2) 考虑到分解重构会引起累积误差，因此分解水平不宜过高。本文用 sym5 小波基对 $\overline{\mathrm{SI}}(t)$ 进行 3 层分解，分别提取 1～3 层的高、低频系数，然后对各层高、低频系数进行重构，得到各频带的信号序列。实际副高面积指数变化可通过上述 1～3 层高频和第 3 层低频的重构信号之和可以精确地获得。

(3) 分别建立 1～3 层高频和第 3 层低频共 4 个频段信号各自的 LS-SVM 预测模型，即利用超前 1、2、3、4、5 候的副高面积指数预测第 1 候、第 3 候、第 5 候的副高面积指数。设 $P$、$T$ 分别为 SVM 模型的预报因子输入和预报结果输出序列。即

$$P = [X(t-1), X(t-2), X(t-3), X(t-4), X(t-5)]$$

$$T = [X(m)], \quad m = 1,3,5$$

则每一时次的训练数据对为：$[X(t-1), X(t-2), X(t-3), X(t-4), X(t-5), X(m)]$，SVM 预测模型可以描述为 $T = \phi(P)$，式中 $\phi$ 为非线性映射。本文构造一个多输入、单输出的 LS-SVM 预测模型。在建立各频段的 LS-SVM 模型时，需要通过大量的试验不断

调试各模型参数。每个模型的建模信号不同，最后确定的模型参数也不同。

(4) 分别将每个频段的独立检验样本、模型参数代入到对应的每个模型，可得各频段的预测信号，重构各频段的预测信号，并对重构后的信号反归一化，即可得预测的副高面积指数。

## 5 结果分析

为检验 WT～LS-SVM 模型的独立预报效果，将 10 年共 730 候的样本分为两部分，第一部分取前 200 候用于模型的建立和测试。在建立模型的过程中采用 $k$-折交叉检验的方法，其基本思想是把 $l$ 个样本点随机地分成 $k$ 个互不相等的子集，即 $k$-折 $S_1, S_2, \cdots\cdots, S_k$。每个折的大小大致相等，共进行 $k$ 次训练与测试，即对 $i=1,2,\cdots\cdots,k$ 进行 $k$ 次迭代，第 $i$ 次迭代的做法是，选择 $S_i$ 为测试集，其余 $S_1,\cdots\cdots,S_{i-1},S_{i+1},\cdots\cdots,S_k$ 的合集为训练集。本文取 $k$=10。第二部分为后 530 候，用于模型的独立检验和预报优化效果的评估。该部分资料不参与模型的建立。

第 1 候和第 3 候各频域分量的 LS-SVM 模型集成结果均比较准确地逼近实际信号（相关系数分别为 0.979 2 和 0.913 1，置信度 $\alpha$ =0.05），在总体趋势上和局部细节上均很好地逼近实际信号。尤其是对副高的几次异常变化的预测均比较准确（图 2、3，为了使信号的局部特征清晰，图中只给出了 218 候的变化曲线，下同）。

第 5 候各频域分量的 LS-SVM 集成结果虽然在细节上和实际信号有较大的出入，但总体趋势上仍能比较准确的逼近实际信号（相关系数为 0.853 7，置信度 $\alpha$ =0.05，图 4）。

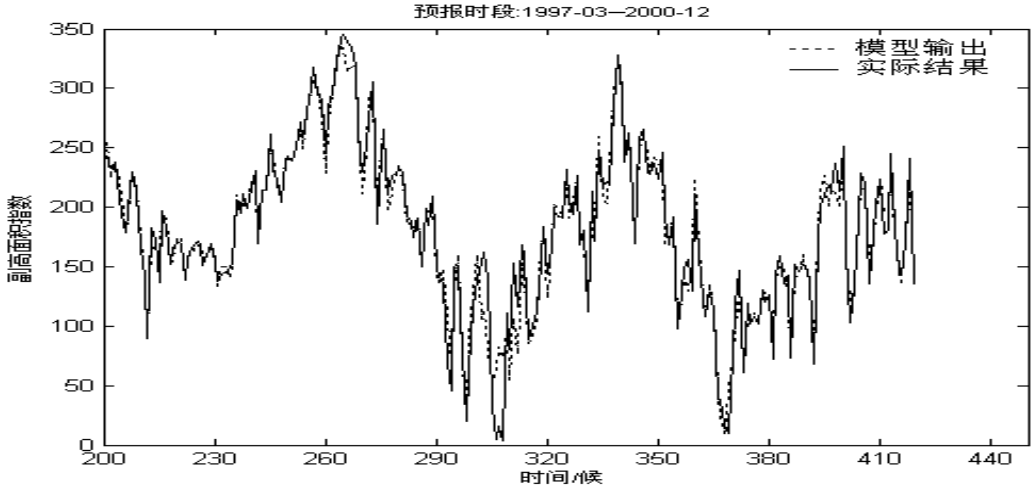


图 2 实际信号（实线）与集成预报（虚线） 预报时效 1 候。

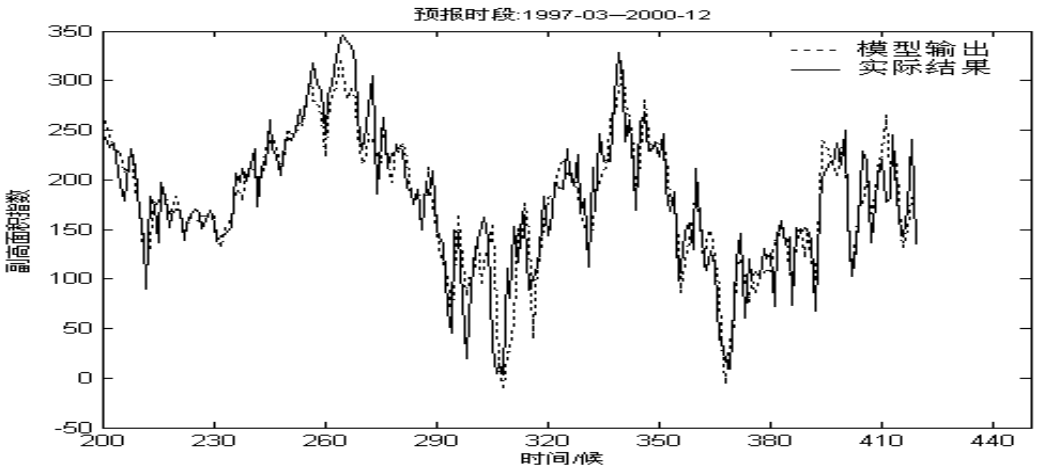


图 3 实际信号（实线）与集成预报（虚线） 预报时效 3 候。

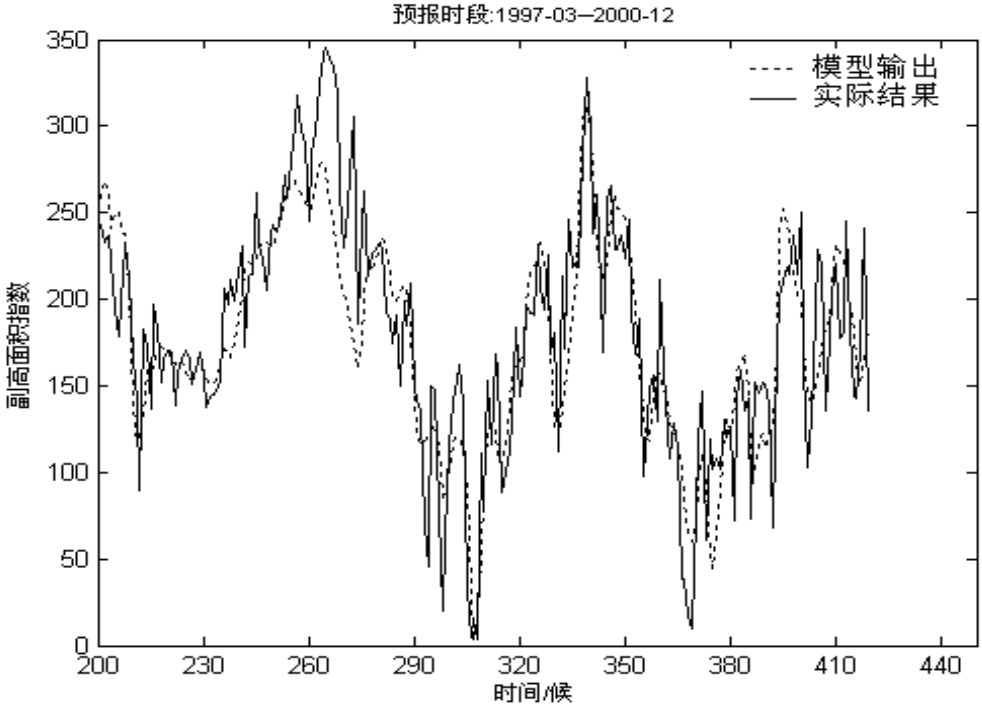


图 4　实际信号（实线）与集成预报（虚线）　　预报时效 5 候。

## 6　不同方法的预报优化效果比较

神经网络和线性回归方法是气象资料分析和要素预测中的常用的方法，为了评估和比较WT～LS-SVM模型的预报效果和技术优势，我们采用同样的数据资料，分别建立副高面积指数的神经网络、线性回归及LS-SVM的预报模型，并与WT～LS-SVM模型的输出结果进行对比。表1为各种预报模型（最小二乘支持向量机、神经网络和线性回归模型）独立检验结果与实际副高指数的相关系数。从表1可看出，WT～LS-SVM模型的独立检验的相关系数均高于神经网络、线性回归模型。上述的对比试验结果表明，本文所采用的小波分解和最小二乘支持向量机相结合的方法对副高预报对象的把握和描述较前面神经网络和线性回归模型更为恰当和准确，表现出较好的预报效果和技术优势。

表 1　不同模型预报结果与实际值相关系数

| 预报模型 | 第 1 候 | 第 3 候 | 第 5 候 |
|---|---|---|---|
| WT～LS-SVM | 0.979 2 | 0.913 1 | 0.853 7 |
| 模糊神经网络 | 0.800 8 | 0.674 9 | 0.626 6 |
| 线性回归 | 0.815 0 | 0.700 4 | 0.645 5 |

## 7　小　　结

本文采用小波分解（WT）和最小二乘支持向量机（LS-SVM）相结合的集成预测方法，用于西太平洋副热带高压的面积指数的预测研究。由于它的变化具有多尺度性和复杂性的特点，首先利用小波分解的方法将副高面积指数分解为相对简单的不同频域的信号，利用 LS-SVM 方法对不同频域的信号分别建立预测模型，最后集成各频域的独立预测结果。为了进一步说明该方法的优势，最后利用同样的资料，分别建立了 LS-SVM、神经网络和线性回归的预测模型，并对每个模型的独立检验结果作了比较分析。试验结果表明，本文采用的小波分析和最小二乘支持向量机相结合的方法在研究和预测西太平洋副热带高压活动等复杂现象时具有明显的优越性。该方法具有简单快捷、泛化能力强、稳定性好等优点，具有广阔的应用前景。

尽管该模型取得了比较满意的预报结果，但是该模型仅利用了副高活动本身的相关信息，即只选取副高前期活动作为预报因子，预报因子不够充分。同时，模型参数的选择缺乏理论依据，需要不断地调试。因此，有必要作更深入研究，更充分地选择预报因子，通过其它优化算法（如模拟退模、遗传算法等）来客观选取确定最小二乘支持向量机

模型的参数，以达到更高、更准确的预报效率。这也　　是我们下一步拟开展的工作。

## 参 考 文 献：

[1] 张韧. 基于前传式网络逼近的太平洋副热带高压活动的诊断预测[J]. 大气科学, 2001, 25(5): 649-660.

[2] ZHANG Ren(张韧), YU Zhihao. Neural Network BP Model Approximation and Prediction of Complicated weather Systems[J]. ACTA Meteorologica Sinica, 2001, 15(1): 105-115.

[3] VAPINK V, GOLOEICH S, SMOLA A. Support vector method for function approximation, regression estimation, and signal processing[M]. Cambridge, MA, MIT Press, 1997: 281-287.

[4] SMOLA A J. Regression estimation with support vector learning machines[D]. Technische University at Myunchen, 1996.

[5] VLADIMIR N Vapnik. 统计学习理论本质[M]. 张学工译. 北京：清华大学出版社，2000: 63-70.

[6] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.

[7] MALLAT S. A theory for multiresolution signal decomposition: the wavelet representation[J]. IEEE Pattern Anal And Mathchine Intell, 1989, 11(7): 674-693.

[8] 胡昌华，张军波，夏军, 等. 基于 Matlab 的系统分析与设计——小波分析[M]. 西安: 西安电子科技大学出版社, 1999: 17-20.

[9] 余丹丹，张韧，董兆俊，等. 近赤道海温影响西太平洋副高的小波多尺度分析[J]. 热带气象学报, 2004, 20(5): 593-599.

[10] 张礼平，陈永义，周筱兰. 支持向量机（SVM）及其在场预测中的应用[J]. 热带气象学报，2006, 22(3): 278-282.

# AREA EXPONENT OF WESTERN PACIFIC SUBTROPICAL HIGH FORECAST MODEL BASED ON WAVELET DECOMPOSITION SUPPORT VECTOR MACHINE

LIU Ke-feng[1, 2], ZHANG Ren[1, 2], YU Peng[1], WANG Yan-lei[1], YU Dan-dan[1]

（1. Institute of Meteorology, PLA University of Science and Technology, Nanjing 211101, China;
2. I Guangzhou nstitute of Tropical and Marine Meteorology, CMA, Guangzhou 510080, China）

**Abstract**：Based on the method of wavelet decomposition and support vector machine, the area index of western pacific subtropical high forecast model was established. By using this method, the area exponent of western pacific subtropical high was decomposed into several relative simple band-pass signals. Then the independent prediction models of decomposed signals with support vector machine were set up, and independent predicted results were integrated. Finally, in order to assess and compare the effectiveness and technical superiority of the paper used, the independent testing results with different models, such as WT～LS-SVM, neural networks and linear regression model were discussed in the same condition. The testing results showed that the model based on support vector machine exhibited its properties of high forecast accuracy, fast training, high generalization capability and easy modeling.

**Key words:** wavelet decomposition; least square support vector machine; subtropical high