杨仲江,蔡波,刘旸. 2013. 利用双隐层 BP 网络进行雷暴潜势预报试验——以太原为例. 气象,39(3):377-382.

利用双隐层 BP 网络进行雷暴潜势预报试验——以太原为例*

杨仲江[1] 蔡 波[1,2] 刘 旸[3]

1 南京信息工程大学大气物理学院,南京 210044

2 沈阳军区空军气象中心,沈阳 110015

3 辽宁省人工影响天气办公室,沈阳 110016

提 要: 利用太原地区探空资料,结合闪电定位资料,采用神经网络法对太原地区雷暴天气进行潜势预报。选取与雷暴发生相关性较好的探空因子作为预报因子,对其进行归一化处理,输出采用两级分类,构建双隐层的 BP 网络,并应用独立样本进行预报检验。结果表明,在相同条件下,与单隐层 BP 网络相比,双隐层 BP 网络显示了其在解决分类问题上的优势;与多元统计回归法相比,双隐层 BP 网络获得更高的雷暴预报 TS 评分及更可靠的结果,显示出神经网络良好的非线性问题处理能力。并且对雷暴预报结果的规律进行了分析与讨论,说明探空因子与雷暴的发生有着密切的联系。

关键词: 神经网络, BP 网络, 雷暴预报, 探空资料

中图分类号: P456 **文献标识码:** A **doi:** 10.7519/j.issn.1000-0526.2013.03.013

Experimental Research on Thunderstorm Forecasting with Double Hidden Layer BP Neural Network: Case Study on Taiyuan

YANG Zhongjiang[1] CAI Bo[1,2] LIU Yang[3]

1 School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing 210044

2 Air Force Meteorological Centre of Shenyang Military Region, Shenyang 110015

3 Weather Modification Office of Liaoning Province, Shenyang 110016

Abstract: A neural network scheme to do a multivariate analysis for forecasting the occurrence of thunderstorm in Taiyuan is presented by using sounding data and lightning location system data. Well correlated sounding factors are selected as the predictors, then all the input factors are normalized, and output data are adopted to two-stage category so that the BP network with double hidden layers has been established and the independent samples can be tested in it. The results indicate that, in the same condition, compared with single hidden layer BP network, the double hidden layer BP network shows its advantage on solving classification problem. Compared with multivariate statistics regression algorithm, the neural network algorithm obtains higher thunderstorm forecasting TS score and more reliable results, showing good nonlinear processing ability in the thunderstorm forecasts based on sounding data. And then the rules of thunderstorm forecast results are analyzed and discussed, showing that sounding factors have a close connection with the occurrence of thunderstorm.

Key words: neural network, BP network, thunderstorm forecast, sounding data

* 公益性行业(气象)科研专项(GYHY200806014)和江苏省高校优势学科建设工程资助项目(PAPD)共同资助
2012 年 2 月 7 日收稿; 2012 年 9 月 5 日收修定稿
第一作者:杨仲江,主要从事雷电研究. Email:mashimaro2974@126.com

引 言

雷暴是一种灾害性天气，其生命史短、空间尺度较小、具有局地性特征，一直是天气预报重难点之一。如何提高雷暴预报的准确率，加强对雷暴天气的分析与研究，对防雷减灾工作至关重要。

目前在国内利用探空因子进行雷暴预报的研究中，多采取多元统计回归的方法。这种方法不可避免的问题就是雷暴发生的非线性与线性回归算法之间的矛盾，因此需要大力发展非线性算法。Agostinon(2005)利用探空资料和闪电定位资料对神经网络进行训练，并对雷暴的发生以及闪电密度进行了预报。赵旭寰等(2009)尝试利用神经网络法对南京地区雷暴天气进行预报。这些研究表明，神经网络预报雷暴发生是可行的。但由于研究重点和地域的不同，网络的构建方式存在一定差异，预报结果判定方式也存在一定可变性。因此，如何解决上述问题，合理构建和使用网络，从而达到更好的预报效果，具有十分重要的意义。

通过对雷暴发生的影响因子进行分析计算，从中发掘有用的预报信息作为模型输入，尝试采用双隐藏层 BP 网络及两级分类的输出，对预报因子归一化处理，试图寻找泛化能力最优的网络来进行雷暴预报的建模研究，以期获得更高的预报准确率和更可靠的预报结果判定方式。

1 资料介绍及处理

1.1 资料选取

选用 2007—2008 年太原站(37.78°N、112.55°E)的探空资料和山西省闪电定位资料。太原站每天进行两次探空，获取探空资料时间为北京时间 08 和 20 时，给出每天两次大气状况。闪电定位资料由山西省气象局提供。为了对网络进行训练及验证，将 2007 年的数据作为训练样本，2008 年的数据作为独立检验样本，采用探空因子作为输入矩阵，雷暴发生与否作为输出矩阵。

1.2 预报量的处理

据统计，太原雷暴天气多发生于夏季，闪电定位系统探测到的地闪有 70%发生在 6—8 月，因此本文将研究时间限定到 6—8 月里。

雷暴的发生是小概率事件，为提高预报量序列中雷暴的发生概率，以太原站为中心，50 km 为半径圆周范围，若探空后 12 h 内，闪电定位网探测到 1 或 1 次以上且电流强度在 10～100 kA 的闪电过程，则认为太原站发生雷暴。

至此，除去缺失资料的样本，网络训练样本对为 179 个，其中有闪电记录的为 84 个，雷暴样本概率为 46.9%。这里将以这 179 个样本对作为训练样本，进行 BP 神经网络的建模研究。

1.3 预报因子的选取

雷暴发生、发展的基本条件有环流条件、水汽条件、不稳定条件及抬升条件。国内外许多学者研究认为，闪电活动与大气不稳定参数存在一定的相关性(郑栋等，2005)。探空因子备选集采用美国怀俄明州立大学提供的探空资料指数产品，共计 20 个。其中对流有效位能 CAPE、虚温计算的 CAPE、对流抑制指数 CIN、虚温计算的 CIN、粗里查森数 BRN、虚温计算的 BRN、500～1000 hPa 厚度 7 个指数由于样本数不全舍弃，虚温计算的 LI 与 LI 对比分析舍弃虚温计算的 LI。至此本文保留了 12 个预报因子，具体定义见表 1。为选择合适的预报因子作为输入，首先分别将这 12 个因子与预报量作相关性分析。

由于对流参数物理量是连续型因子，雷暴发生与否为 0，1 变量，所以不直接求出其相关系数，而是求出其点双序列相关系数。具体而言，当 X 是连续型因子，Y 是 0，1 变量，它们之间的相关系数为：

$$r=\frac{\overline{x}_{(1)}-\overline{x}}{S_x}\left(\frac{P}{1-P}\right)^{\frac{1}{2}} \tag{1}$$

式中，$\overline{x}$ 为因子 x 的平均值，$\overline{x}_{(1)}$ 为在 $y=1$ 时 x 的平均值，P 为事件 $y=1$ 出现的频率，S_x 为因子的样本标准差(刘震钊等，2010)。

本文选取相关系数在 0.3 以上的 7 个指数作为预报因子，分别是 CT 指数、K 指数、抬升指数、累积可降水量、沙氏指数、强天气威胁指数和 TT 指数。

其中 K 指数在考虑了大气层结不稳定的同时，还考虑了对流层中层的水汽条件。CT 指数、TT 指数表征了雷暴发生前的大气层结的不稳定条件。抬升指数 LI 表征大气不稳定度，当 $LI<0$ 时，大气层结不稳定，且负值越大，不稳定程度越大，反之，则表示大气层是稳定的。沙氏指数SI和强天气威胁指

表 1　预报因子及其定义

Table 1　The predictors and their definition

预报因子名称	缩写	定义及物理含义	相关系数
CT 指数	CT	$CT = T_{d-850} - T_{500}$	0.3942
K 指数	KI	$K = (T_{850} - T_{500}) + T_{d-850} - (T_{700} - T_{d-700})$	0.5258
抬升指数	LI	$LI = T_{500} - T_{i-500}$	−0.4831
累积可降水量	PWFES	对伴有强降水的雷暴预报有作用	0.4043
沙氏指数	SI	层结不稳定参数。一般 $SI<0$ 时,大气层结不稳定,且负值越大,不稳定程度越大;反之,则表示大气层结是稳定的	−0.4577
强天气威胁指数	SWI	强对流天气过程的诊断分析中常用的物理量	0.3277
TT 指数	TT	$TT = VT + CT$	0.4189
混合层平均混合比	MMLMR	与大气层结稳定度有关	0.2500
混合层平均潜热	MMLPT	与大气层结稳定度有关	0.1511
抬升凝高度层气压	PLCL	与大气层结稳定度有关	0.0596
抬升凝结高度层温度	TLCL	与大气层结稳定度有关	0.2019
VT 指数	VT	$VT = T_{850} - T_{500}$	0.1956

数 SWI 属于稳定度指标。累积可降水量 PWFES 代表水汽指标。

2　神经网络的构建

人工神经网络(Artificial Neural Network, ANN)是理论化的人脑神经网络的数学模型,是基于模仿大脑神经网络结构和功能而建立的一种信息处理系统。它实际上是由大量简单原件相互连接而成的复杂网络,具有高度的非线性,能够进行复杂的逻辑操作和非线性关系实现的系统(袁曾任,1999)。

在人工神经网络的实际应用中,应用最广泛的就是 BP 网络模型,它体现了人工神经网络最精华的部分。近年来 BP 神经网络也越来越多的应用于大气科学领域(刘旸等,2011;官莉等,2010;张雪慧等,2009;农孟松等,2011;程炳岩等,2011),这种网络模型相对于其他模型而言,通用性好,且较为成熟。相对传统的数理统计方法而言,BP 神经网络可以求解非线性问题,同样对样本大小的要求也可以相对少得多,BP 网络模型示意图,如图 1 所示。

BP网络可有效地用于复杂的非线性函数的逼近,一个 3 层的前馈网络能够实现任意精度的连续函数映射(丛爽,2003),两个隐层的神经网络则可以解决各种分类问题(金龙,2004)。因此本文尝试构建两个隐藏层的 BP 网络。

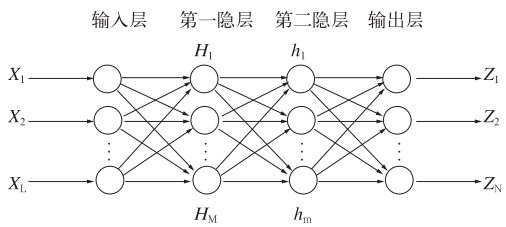

图 1　双隐层 BP 网络模型

Fig. 1　The double hidden layer BP network model

2.1　输入输出数据的处理

训练样本集是影响 BP 网络性能优劣关键。神经网络是以样本在事件中的统计几率来进行训练和预测的。归一化是归纳统一样本的统计分布性,可以简化计算,缩小量值,加快网络的收敛。归一化在[0,1]之间是统计的概率分布,归一化在[−1,+1]之间是统计的坐标分布。兼顾 sigmod 函数的定义,这里采取式(2)对输入样本进行[0.1,0.9]之间的归一化:

$$\boldsymbol{X} = 0.1 + 0.8[(\boldsymbol{P} - \min\boldsymbol{P})/(\max\boldsymbol{P} - \min\boldsymbol{P})] \tag{2}$$

这里,$\boldsymbol{P}$ 表示归一化前的输入数据,$\min\boldsymbol{P}$ 表示矩阵 $\boldsymbol{P}$ 的最小值,$\max\boldsymbol{P}$ 表示矩阵 $\boldsymbol{P}$ 的最大值,$\boldsymbol{X}$ 表示归一化之后的输入矩阵。

需要说明的是,训练前,2007 年的数据进行归一化;仿真前,2008 年的样本要采取与 2007 年同样的设置来归一化。

对于输出矩阵,在 12 h 内发生一次雷暴即认为该时段有雷暴发生。输出矩阵按雷暴的发生与否分成两类,并考虑到 sigmod 函数的输出范围,无雷暴设计成[0.1,0.9],有雷暴设计成[0.9,0.1]。

2.2　学习算法

在选择算法对网络进行训练的时候,在网络参

数很多，需要考虑存储容量问题时，选择共轭梯度法。经试验，这里选择 Scaled 共轭梯度法 trainscg。

2.3 传递函数

各层神经元的传递函数根据需要，不同层内采取不同的传递函数组合。并且一般常采用 S 形函数和线性函数组合。S 形函数能将跨度很大的数值压缩到一个很小的固定范围内。输出层常采用线性函数，这样整个网络的输出可以取实数域内任何值。双隐层的网络结构，需要 3 层传递函数，选取 tansig 和 purelin 的结合，在试验中，最终确定 tansig-tansig-purelin 组合。这两种函数的表达式如下：

$$\boldsymbol{A} = f(\boldsymbol{W} \cdot \boldsymbol{P} + \boldsymbol{b}) = \frac{1}{1 + \mathrm{e}^{-(\boldsymbol{W} \cdot \boldsymbol{P} + \boldsymbol{b})}} \tag{3}$$

$$\boldsymbol{A} = f(\boldsymbol{W} \cdot \boldsymbol{P} + \boldsymbol{b}) = \boldsymbol{W} \cdot \boldsymbol{P} + \boldsymbol{b} \tag{4}$$

其中，$\boldsymbol{P}$ 表示输入矩阵，$\boldsymbol{W}$ 表示权值矩阵，$\boldsymbol{b}$ 表示偏置矩阵，$\boldsymbol{A}$ 表示输出。

2.4 过拟合问题的处理

怎样寻求泛化能力最大的网络，是网络构建的一个难题。网络的泛化性能是指神经网络对训练样本以外的新样本的适应能力。在网络的训练过程中经常会出现过拟合现象，即在网络训练过程中，从某一次训练开始，随着网络的训练能力提高，仿真的能力反而下降。解决这个问题的一种办法是：在数据输入中，给训练的数据分类，分为正常训练用、变量数据和测试数据，3 种数据分别占用全部样本的 0.6、0.2和 0.2。设定参数 maxfail＝10（最大失败次数），在网络训练的过程中，如果从某一步开始，变量数据的误差不降反升，则 Validation Checks 开始计数，当计数到 10，则认为网络陷入过拟合，停止训练。

2.5 隐节点数的选取

目前隐节点的选取没有一个固定的参照，主要靠经验试凑。

$$h = \sqrt{n + m + a} \tag{5}$$

其中，h 为隐节点数，n 为输入节点数，m 为输出节点数，a 为 1～10 之间的常数。

隐节点数 h 与输入节点 n 的关系为：

$$h = \log_2 n \tag{6}$$

在神经网络研究中，高大启（1998）用曲面对隐节点数的规律进行拟合，由最小二乘法得到拟合公式为：

$$h = \sqrt{0.43mn + 0.12m^2 + 2.54n + 0.77m + 0.35} + 0.51 \tag{7}$$

隐节点数太少会造成信息不足，造成网络训练达不到要求；太多会造成浪费，使训练时间加长。

由于这些公式只是大致给出了隐节点的取值范围，而不同的研究问题之间有很大的差别，所以仅根据上述公式来确定隐节点数不是十分合适。这里将其作为参考，尝试多种隐节点数组合，根据网络训练的 MSE，最终确定两层隐节点数分别为 3 和 9（详见表 2）。

表 2 隐节点的选取

Table 2 Selecting hidden nodes

	第一层隐节点数	第二层隐节点数	MSE
1	3	11	0.0807
2	3	5	0.0859
3	3	9	0.0703

3 预报模型的检验分析

3.1 双隐层 BP 网络预报结果

选用 2008 年 6—8 月的 167 个样本作为独立样本进行预报检验，其中雷暴样本 69 个，非雷暴样本 98 个，预报结果见表 3。预报结果令人满意，表明此神经网络模型适用于太原市雷暴预报。

表 3 2008 年 6—8 月 167 个独立样本预报结果

Table 3 Forecasting results with the 167 independent samples from June to August 2008

	实况有	实况无	预报结果	
预报有雷暴	a＝雷暴	b＝空报	a＝60	c＝9
预报无雷暴	c＝漏报	d＝无雷暴	b＝17	d＝81

采用 TS 评分和准确率来检验预报结果：

$$TS = \frac{a}{a + b + c} = \frac{60}{60 + 17 + 9} = 69.77\% \tag{8}$$

$$\text{准确率} = \frac{a + d}{a + b + c + d} = \frac{60 + 81}{167} = 84.43\% \tag{9}$$

3.2 其他几种算法的预报结果

为对比分析双隐层 BP 网络在预报雷暴问题上的优势，使用同样的训练样本和独立检验样本，分别采用单隐层 BP 网络、多元线性回归算法及最优线

性回归算法进行预报。

(1) 采用单隐层 BP 网络算法预报结果:在 167 个独立检验样本中,雷暴样本报对 62 次,非雷暴样本报对 68 次,漏报 7 次,空报 30 次。其 TS 评分为 62.63%,准确率为 77.84%。

(2) 采用多元线性回归算法进行预报建模,得到如下回归方程:

$$\begin{aligned} Y = &-3.6870 - 0.0284CT + 0.0089KI - \\ &0.0196LI + 0.0268PWFES + \\ &0.0925SI + 0.0011SWI + \\ &0.0765TT \end{aligned} \tag{10}$$

雷暴样本报对 56 次,非雷暴样本报对 76 次,漏报 13 次,空报 22 次。其 TS 评分为 61.54%,准确率为 79.04%。

(3) 采用最优线性回归算法进行预报建模试验。采用一种逐步回归方法求得的局部最优子集来取代全局最优子集,经过计算,选取了 5 个因子。因子选取方法取自李东风等(2008)的研究结果,得到如下回归方程:

$$\begin{aligned} Y = &\ 6.9764 + 0.0059CT + 0.0041KI - \\ &0.0653LI - 0.0297MMLPT - \\ &0.0207SWI \end{aligned} \tag{11}$$

雷暴样本报对 55 次,非雷暴样本报对 78 次,漏报 14 次,空报 20 次。其 TS 评分为 61.8%,准确率为 79.64%。

3.3　预报结果对比分析

为了说明双隐层 BP 网络预报在雷暴上的优势,以雷暴样本为例,将网络的两级输出绘图,如图 2 所示。若实心圆点数值大于空心圆点,即该样本是雷暴的可能性大于非雷暴,则认为是雷暴样本;反之是非雷暴样本。相较于线性回归算法,这种方法的优势在于不需要设定阈值,结果更可靠。

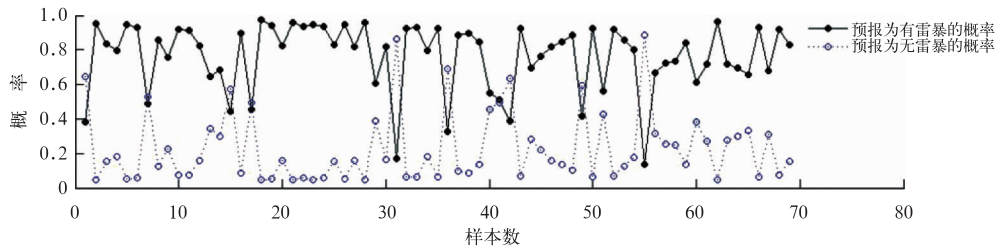

图 2　双隐层网络实际输出数值特征

Fig. 2　Statistical characters of actual output from double hidden layer BP network

图 3 给出了双隐层 BP 网络与多元线性回归算法 69 个雷暴样本的输出对比图。由图 3 可以看出,多元线性回归法的概率值比较分散,双隐层 BP 网络法的输出结果大部分都集中在[0.8,1]区间内,更加接近真值。

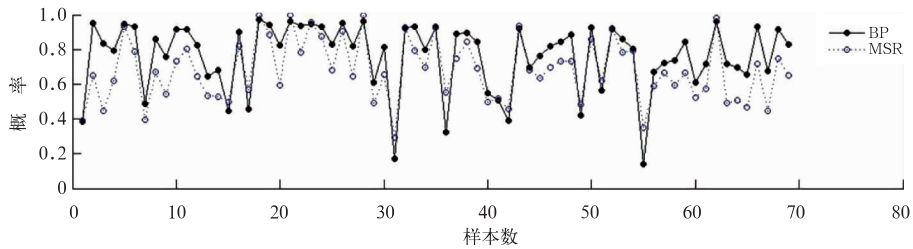

图 3　BP 网络法与多元线性回归法预报结果对比

Fig. 3　Comparison of forecasting results between BP network and multivariate statistics regression method

为对比双隐层 BP 网络在解决分类问题上的优势,图 4 给出了单隐层 BP 网络的输出,同样以雷暴样本为例。可以看出,单隐层 BP 网络的预报结果两级化差异不如双隐层明显,这将导致预报难度大的样本出错的可能性更大,从数据上也验证了这一点,单隐层BP网络空报次数过多,导致TS评分降

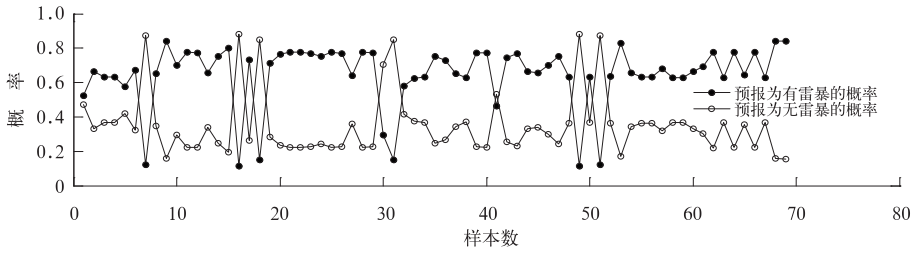

图 4 单隐层 BP 网络实际输出值特征

Fig. 4 Statistical characters of actual output from single hidden layer BP network

低。由此可见,在解决分类问题上,双隐层 BP 网络相对于单隐层 BP 网络有着一定的优势。

4 结 论

(1) 在预报结果判定方面,网络采用单输出时,结果往往需要设定阈值来判定,而阈值的设置与雷暴的时空分布有关,数据预报结果可靠性不高。本文网络的输出采用二级分化,通过两级结果比较,可以直接判定雷暴发生与否,不需要设定阈值,可信度高。

(2) 在以往的研究中,雷暴预报多采取单隐层的 BP 网络。两个隐层的神经网络则可以解决各种分类问题,结果表明,双隐层 BP 网络在解决分类问题上有一定的优势。

(3) 为了寻找泛化能力最高的网络,对网络的输入进行了归一化处理,加快了网络训练速度,提高了网络的精度。针对训练数据进行分类,一定程度上避免了过拟合问题。传递函数及隐节点数的选取均尝试多种不同的组合,根据网络训练 MSE 最终确定最优组合。最后构建了针对本文研究内容的泛化能力较高的网络。

(4) 利用多元线性回归法和神经网络法进行雷暴预报的对比分析,采用相同的训练样本和独立检验样本,结果表明,与多元线性回归法相比,神经网络法在当地的雷暴预报中获得了更高的 TS 评分。

(5) 通过多组网络的预报结果,分析了雷暴预报结果的规律,表明了探空因子的变化与雷暴的发生有着密切的联系。

参考文献

程炳岩,郭渠,张一,等. 2011. 三峡库区高温气候特征及其预测试验. 气象,37(12):1544-1552.

丛爽. 2003. MATLAB 工具箱的神经网络理论与应用. 合肥:中国科学技术大学出版社,35.

高大启. 1998. 有教师的线性基本函数前向三层神精网络格构研究. 计算机学报,21(1):80-86.

官莉,刘旸,张雪慧. 2010. 人工神经网络算法在红外高光谱资料反演大气温度廓线中的应用. 大气科学学报,33(3):341-346.

金龙. 2004. 神经网络气象预报建模理论方法与应用. 北京:气象出版社.

李东风,郑忠国. 2008. 最优线性回归的计算方法. 数理统计与管理,27(1):87-95.

刘宸钊,卓伟,裴军林. 2010. 基于对流参数的雷暴预报方法研究. 高原山地气象研究,30(2):22-25.

刘旸,官莉. 2011. 人工神经网络法反演晴空大气湿度廓线的研究. 气象,37(3):318-324.

农孟松,黄海洪,孙崇智,等. 2011. 基于主分量神经网络的降水集成预报方法研究. 气象,37(3):352-355.

袁曾任. 1999. 人工神经网络及其应用. 北京:清华大学出版社.

张雪慧,官莉,王振会,等. 2009. 利用人工神经网络方法反演大气温度廓线. 气象,35(11):137-142.

赵旭寰,王振会,肖稳安,等. 2009. 神经网络在雷暴预报中的应用初步研究. 热带气象学报,25(3):357-360.

郑栋,张义军,吕伟涛,等. 2005. 大气不稳定度参数与闪电活动的预报. 高原气象,24(2):196-203.

Agostino M. 2005. Sounding-derived indices for neural network based short-term thunderstorm and rainfall forecasts. Atmospheric Research,83(3):349-365.