引用格式：方颖,李连发.基于机器学习的高精度高分辨率气象因子时空估计[J].地球信息科学学报,2019,21(6):799-813. [Fang Y, Li L F. Spatiotemporal estimation of high-accuracy and high-resolution meteorological parameters based on machine learning[J]. Journal of Geo-information Science, 2019,21(6):799-813.] **DOI:**10.12082/dqxxkx.2019.190014

基于机器学习的高精度高分辨率气象因子时空估计

方　颖[1,2],李连发[1,2*]

1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101；2. 中国科学院大学,北京 100049

Spatiotemporal Estimation of High-Accuracy and High-Resolution Meteorological Parameters based on Machine Learning

FANG Ying[1,2], LI Lianfa[1,2*]

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science, Beijing 100101, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The meteorological stations are sparsely distributed across Mainland China. In terms of generating high-resolution surfaces of meteorological parameters, the estimation accuracy of existing models is limited for air temperature, and is poor for relative humidity and wind speed (few studies reported). With the measurement data of 824 monitoring stations covering the mainland of China in 2015, this study compared the typical Generalized Additive Model (GAM) and autoencoder-based residual neural network (here after, residual network for short) in terms of predicting three meteorological parameters, i.e. air temperature, relative humidity, and wind speed. The performances of the two models were evaluated through 10-fold cross-validation. Basic variables including latitude, longitude, elevation, and the day of the year are used in the air temperature models. In addition to the basic variables, the relative humidity models use air temperature and ozone concentration as covariates, and the wind speed models use wind speed coarse-resolution reanalysis data as covariates. In our spatiotemporal models, spatial coordinates capture the spatial variation and time index of the day captures the time variation. Compared with GAM, residual network significantly improved the prediction accuracy: on average, CV (Cross Validation) $R^2$of the three meteorological factors increased by 0.21, CV RMSE decreased by 37%, and the relative humidity model improved the most (CV R^2: 0.85 vs. 0.52, CV RMSE: 7.53% vs. 13.59%). With incorporation of the monthly index in the relative humidity models, the accuracy was greatly improved, indicating that the different levels of time factors are important for the relative humidity models. Furthermore, we also discussed the effectiveness and limitations of coarse resolution reanalysis data and nearest neighbor values as covariates. This study shows that the residual network model can greatly improve the accuracy of national high spatial (1 km) and temporal (daily) resolution meteorological data as opposed to traditional GAMs. Our findings provide implications for high-accuracy and high-resolution mapping of meteorological parameters in China.

收稿日期：2019-01-19；**修回日期**：2019-03-04.

基金项目：国家自然科学基金项目(41471376、41871351)；中国科学院先导研究项目 (XDA19040501)。[**Foundation items:** National Natural Science Foundation of China, No.41471376, 41871351; Priority Research Program of the Chinese Academy of Science, No.XDA19040501.]

作者简介：方　颖(1995-),女,安徽宣城人,硕士生,研究方向为空间数据分析。E-mail: fangying@lreis.ac.cn

***通讯作者**：李连发(1978-),男,贵州黔西人,博士,副研究员,硕士生导师,研究方向为空间数据分析、空间数据挖掘、风险分析。E-mail: lilf@lreis.ac.cn

Key words: meteorological factors; machine learning; residual autoencoder; Mainland China; GAM; deep learning; high resolution

***Corresponding author:** LI Lianfa, E-mail: lilf@lreis.ac.cn

摘要:气象变量常作为重要的影响因子出现在环境污染、疾病健康和农业等领域,而高分辨率的气象资料可作为众多研究的基础数据,对推进相关研究的发展意义重大。本文以中国大陆为研究区域,利用2015年824个气象站点的气温、相对湿度和风速3套数据,结合不同的解释变量组合,分别构建了各自的GAM和残差自编码器神经网络(简称残差网络)模型,以10倍交叉验证判断模型是否过拟合。研究结果表明:① GAM和残差网络方法都不存在过拟合问题,同GAM相比,残差网络显著提高了模型预测的精度(3个气象因素的交叉验证CV R^2平均提高了0.21,CV *RMSE*平均降低了37%),其中相对湿度模型的提升幅度最大(CV R^2:0.85 vs. 0.52,CV *RMSE*:7.53% vs. 13.59%);② 残差模型的结果较普通克里格插值结果和再分析资料更接近站点观测数据,表明残差网络可作为高分辨率气象数据研制的可靠方法。此外,研究还发现在相对湿度模型中加入臭氧浓度和气温、在风速模型中加入GLDAS风速再分析资料,可提升模型的性能。

关键词:气象因素;机器学习;残差自编码;中国大陆;GAM;深度学习;高分辨率

1 引言

气象是影响人们生产生活的基本变量,在环境空气污染领域,气温、相对湿度和风速等气象因子,影响臭氧、$PM_{2.5}$和氮氧化物等污染物浓度的时空分布,是重要的预测因子之一[1-3];在农业领域,农作物的产量、生物物种的发育和繁殖、农业灾害同气象环境密不可分,如孕穗期空气相对湿度升高会导致籽粒产量显著降低[4],较高湿度能促进苹小卷叶蛾(苹果的害虫之一)生长发育,提高其繁殖力[5],气温升高加剧了干旱胁迫对宁夏枸杞光合作用的抑制作用[6];在环境健康方面,气象同某些疾病紧密相关,柏延臣等[7]发现手足口病在中国大陆的分布具有空间异质性,这与气候和社会经济变量有关;在新能源领域,不同于气温和相对湿度,风速还是一种可提供电力的重要能源,通过对风速的研究可以探知风能资源的分布[8-9]。在实际应用中,高精度高分辨率的气象要素的预测是重要的辅助决策信息。

中国的地面气象监测站点稀疏且分布不均匀,2015年全国一共只有824个监测站点,东部沿海及中部地区的站点远远多于西北部地区。如何从有限的监测站点处的气象数据可靠地估测无监测站点处的气象值对许多应用而言都是极为重要的。在本研究涉及的3个气象变量中,对气温进行栅格化的研究是最多的。直接插值法简单且常见,如距离权重反比(Inverse Distance Weighting,IDW)[10-11]、普通克里格法(Origin Kriging,OK)[12-13]、协同克里金法(Co-Kriging,CK)[14-15]、Spline-样条插值法[16-17]等,插值结果在反映小尺度的气候变化规律时细节信息有限[18]。为补偿稀疏覆盖造成的台站偏差,王劲峰等[19-21]使用基于BSHADE(Biased Sentinel Hospitals Areal Disease Estimation)的点插值方法,得到了误差方差最小的无偏估计。除直接插值外,还有趋势面法[22]、多元回归[23-24]等,将宏观地理因子作为参数纳入到气温模型构建中,可以有效地提高大尺度范围内的模拟精度[24],考虑与气温相关的因子来提高气温分布微观细节的模拟精度是未来的发展趋势[25],如将高程和NDVI信息加入到模型中[26-27]。此外,使用模式进行气温数据的栅格化也在研究中[28]。还有少量的研究使用了人工神经网络[29]、地理权重回归[30]、广义线性回归[31]、贝叶斯最大熵模型[32]来估计气温。虽然不乏对气温栅格化的研究,也有一些学者关心长时序范围内的风速变化[33],但是至今为止对相对湿度和风速进行栅格化的研究依然十分稀少。针对现有气象监测站点的稀疏性、估计方法在精度方面的限制、在风速及相对湿度研究的匮乏及应用的重要性,本文首先对气温、相对湿度和风速3个气象要素以及它们的解释变量做统计分析,并且计算皮尔逊相关系数进行探索性分析,结合不同的解释变量组合对这3个气象因子分别构建GAM和残差网络模型,根据模型的结果确定每个气象模型的协变量,比较2种模型生成的气象分布图之间的区别,并将残差网络模型的结果同普通克里格插值结果和再分析资料进行比较以验证其可靠性。

2 研究区概况、数据源与预处理

2.1 研究区概况

研究区域为中国的大陆,不包括台湾省和中国

海域上的陆地，覆盖范围为73°27′E−135°06′E，18°11′N−53°33′N，研究区占地面积9 457 700 km²。中国气候类型多样。东半部具有大范围的季风气候，冬季盛行大陆季风，寒冷干燥；夏季盛行海洋季风，湿热多雨。青藏高原海拔高，面积大，形成独特的高寒气候。西北地区则因僻处内陆，为海洋季风势力所不及，具有西风带内陆干旱气候。图1是研究区边界、高程的气象站点。

2.2 数据源

研究使用的数据包括位置信息（经度、纬度）和一年中的第几天（Day of Year，DOY），既可用于确定位置和时间信息，也可解释时空变异，是气象模型估计的基础地理变量；气温的空间化借助地形因子（高程）提升模型性能，本研究也将高程作为3个气象模型中的协变量。根据领域知识和探索性分析可知，相对湿度同臭氧浓度[34]、气温和月份具有一定的相关性，风速同网格化的风速再分析资料具有较高的相关性，因此分别在2个模型中另加入相应的协变量。

（1）地面气象站点数据

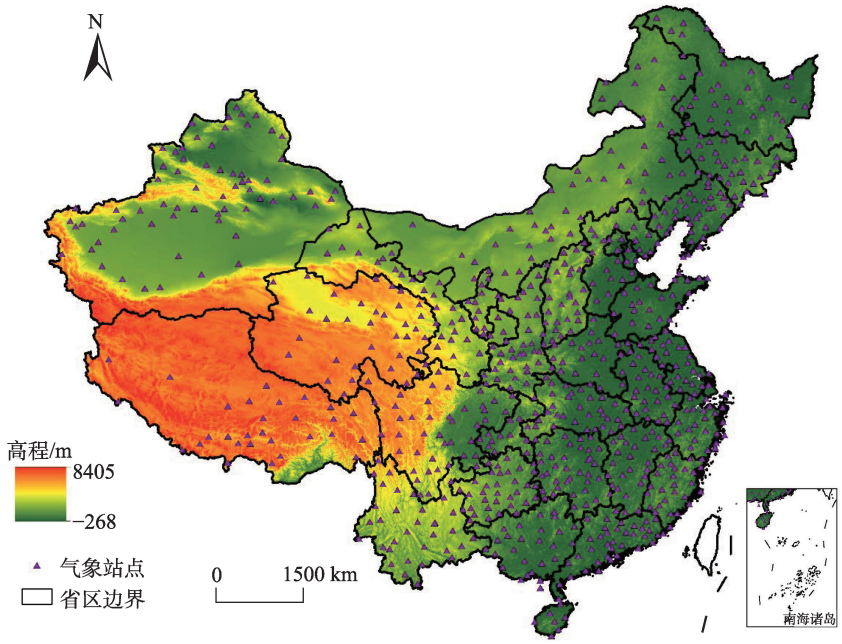
地面气象站点数据来自国家气象科学数据共享平台发布的中国地面气候资料日值数据集①（V 3.0），包含中国824个国家级基准、基本气象站，研究使用了2015年的气温、相对湿度和风速数据。经数据筛选后只保留质量控制为“数据正确”的值，2015年气温、相对湿度和风速数据中符合条件的气象站点数目分别为818、818和770。

（2）SRTM

高程数据来自资源环境数据云平台对航天飞机雷达地形测绘任务（Shuttle Radar Topography Mission，SRTM）进行重采样后的500 m空间分辨率的DEM数据②。SRTM是美国太空总署（NASA）和国防部国家测绘局（NIMA）以及德国与意大利航天机构共同合作完成联合测量，在2000年2月采集数据，2003年公开发布的DEM，覆盖地球80%以上的陆地表面，可免费获取。

（3）GLDAS

在风速模型的构建中使用了来自全球陆面数据同化系统(Global Land Data Assimilation System，GLDAS)的3 h平均风速数据。GLDAS是由美国哥达德空间飞行中心（GSFC）和美国国家环境预报中心（NCEP）联合发展的全球高分辨率的陆面模



注：由于数据获取困难，本次研究不包括香港、台湾和澳门。该图基于国家测绘地理信息局标准地图服务网站下载的审图号为GS(2016)1570号的标准地图制作，底图无修改。

图1　研究区边界、高程和气象站点

Fig. 1　Boundary, elevation and meteorological sites of the study area

① http://data.cma.cn/。

② http://www.resdc.cn/。

拟系统,可提供全球范围1979年至今的陆面资料,空间分辨率为0.25°×0.25°,时间间隔为3 h。GLDAS数据集可以从NASA戈达德地球科学数据和信息服务中心(GES DISC)免费获取。此外,GLDAS也提供了风速和气温的再分析资料用于模型结果的验证。

(4)GEOS-FP

在构建相对湿度的模型时用到了臭氧浓度变量,数据来自于戈达德地球观测系统-前向处理(Goddard Earth Observing System-Forward Processing,GEOS-FP)。GEOS-FP是由GEOS数据同化系统(Data Assimilation System,DAS)产生的最新的GEOS-5气象数据产品。GEOS-FP数据可以覆盖整个中国,数据的空间分辨率为0.25°(纬度)×0.3125°(经度),时间间隔为3 h(ftp://rain.ucis.dal.ca/ctm/GEOS_0.25x0.3125_CH.d/GEOS_FP)。

(5)NCEP/NCAR

NCEP/NCAR是美国国家环境预报中心(National Centers for Environmental Prediction,NCEP)和美国国家大气研究中心(National Centers for Atmospheric Research,NCAR)的联合产品,是一个不断更新的全球网格数据集,代表了地球大气的状态。研究使用其包含的相对湿度再分析资料验证相对湿度模型结果的可靠性。

上述协变量来源等信息的说明见表1。

2.3 数据预处理

研究对3个气象因子分别构建模型,因此有3套数据。气温、相对湿度和风速模型都以了经度、纬度、高程和DOY作为协变量。另外,相对湿度模型的协变量还有日均气温、日均臭氧浓度和最近邻值(最近邻值指的是距离当前点最近的一个站点的日均相对湿度值,如距离点a最近的站点是站点b,那么点a在某一天的最近邻值就是站点b在此天的值),其中日均气温来自本研究中产生的气温日均数据,日均臭氧浓度来自GEOS-FP,最近邻值利用python的SciPy包计算得到(计算方法是k-d树,k-d树是每个节点都为k维点的二叉树,可应用于最近邻搜索,可调用python的scipy.spatial.cKDTree进行计算,参数是2个包含点位置信息的数据集合(在本研究中,模型训练阶段都是实际的站点数据集,模型估计阶段是实际的站点数据集和用以估计的格网点数据集));风速模型还以GLDAS风速作为协变量。GEOS-FP臭氧浓度和GLDAS风速的时间间隔都为3 h,研究对一天中的8个3 h进行平均得到日均值,影像数据在应用之前需要先在R中重采样到结果气象栅格格网。

3 研究方法

3.1 相关性分析

构建高精度、高性能的模型离不开与被解释变量显著相关的解释变量。本研究计算了被解释变量与解释变量之间的皮尔逊相关系数,用以初步判断解释变量可能的贡献。

3.2 广义加性模型(GAM)

GAM是一种广义非线性模型(GLM),模型中预测因子线性依赖于某些预测变量的未知光滑函数。此模型关联单一的响应变量Y和一些预测变量x_i,Y指定为一个指数族分布(如正态分布、二项式分布和泊松分布),连接函数g(如身份函数和log函数)通过式(1)关联Y的期望和这些预测变量:

$$g(E(Y))=\beta_0+f_1(x_1)+f_2(x_2)+\cdots+f_m(x_m) \quad (1)$$

式中:f_i既可以是指定参数形式的函数,也可以是非指定参数或者半参数函数,简称为光滑函数,通过

表1 协变量来源

Tab. 1 Sources of the covariables

数据	来源	空间分辨率	时间间隔
经度	中国地面气候资料日值数据集	-	1 d
纬度	中国地面气候资料日值数据集	-	1 d
DOY	中国地面气候资料日值数据集	-	1 d
高程	SRTM	500 m	-
风速再分析资料	GLDAS	0.25°	3 h
气温再分析资料	GLDAS	0.25°	3 h
臭氧浓度	GEOS-FP	0.25°×0.31°	3 h
相对湿度再分析资料	NCEP/NCAR	2.5°	1 d

非参数手段估计。它的非参数形式使得模型非常灵活，揭示出模型的非线性效应。使用R中的mgcv包对气温、相对湿度和风速构建GAM模型，非线性GAM模型的公式如下：

$$Tem_{s,t}=s(x,y)+s(ele_s)+as.factor(DOY) \quad (2)$$

$$RH_{s,t}=s(x,y)+s(ele_s)+s(tem_{s,t})+s(ozone_{s,t})+as.factor(month)+as.factor(DOY) \quad (3)$$

$$Wind_{s,t}=s(x,y)+s(ele_s)+s(im\_value_{s,t})+as.factor(DOY) \quad (4)$$

式中：s代表时间；t代表位置；x,y代表经纬度信息；*ele*代表高程值；*DOY*代表一年中的第几天；*month*代表月份；*tem*代表气温；*RH*代表相对湿度；*Wind*代表风速；*ozone*代表GEOS-FP臭氧浓度值；*im_value*代表GLDAS风速再分析资料值。

3.3 深度学习模型

（1）全连接前馈神经网络

神经网络是指一系列受生物学和神经学启发的数学模型。全连接前馈神经网络是一种神经网络，这种网络每一层的神经元和下一层的神经元之间都相互连接。针对不同的问题，使用不同的网络结构。一个网络结构的确定需要指定网络的层数、每一层的神经元数目和激活函数，网络结构的确定相当于模型的确定，与多数的机器学习过程一样，确定了模型之后需要指定损失函数作为训练目标，指定优化算法作为训练方法。

（2）自编码器模型

自编码器是前馈网络的一种，它是一种具有对称结构的神经网络，输入变量和输出变量的数目相同，可以分为编码和解码2个阶段。

假设一个d维空间的输入、输出x，权重矩阵w，偏置向量b，参数集合θ，网络层数的索引L，将输入映射到输出，本文可以得到以下：

$$\theta_{W,b}(x):R^d \rightarrow R^d \quad (5)$$

$$\theta_{W,b}(x)=f(W^{(L)}f\left(\cdots f\left(W^{(1)}x+b\right)\cdots\right)+b^{(L)}) \quad (6)$$

参数W,b由最小化训练数据上的x和$\hat{x}$之间的L损失得到：

$$L=\frac{1}{2}\|x-\hat{x}\|^2=\frac{1}{2}\left\|x-\theta_{W,b}(x)\right\|^2 \quad (7)$$

自编码器提供了一个平衡的网络拓扑结构，在编码到解码的映射过程中实现了同主成分分析类似的变量转换功能，而同输入相同的输出类型类似于加入了正则化因素，可有效地防止过拟合。

（3）基于自编码器的残差网络模型

残差深度网络模型是本文提出的主要用于多元回归预测的深度学习模型[35]。该模型在自编码器的基础上衍生的，通过加入残差连接（Residual Connection）实现了误差信息的高速传递，有效地弥补了常规深度网络随着网络层数的加深精度的下降问题。

自编码器是实现残差深度网络的基础部件。自编码器具有编码和解码层的镜像网络。对编码层，每个隐藏层可能有不同数量的节点，这些节点可确保网络变量的变化，起到网络维度的压缩或调整；对解码层，每层都对应了编码层（具有相同的层结点数），二者之间对应而实现了残差连接。对于残差回归网络，自编码器是一种自然的选择，因为残差连接通常要求2个（浅层和深层）实现它们之间的直接连接的节点数量相同。在本文前期的研究中残差深度网络极大地提高了预测精度及收敛速度，更具体的细节参见前期研究结果[35]。

本研究中风速模型和气温模型具有相同的网络结构：5层神经网络，每一层的神经元数目分别为196、128、96、64和32；相对湿度模型具有5层神经网络，每一层的神经元数目分别为256、128、96、64和32。3个模型都使用修正线性单元（Rectified Linear Unit, ReLU）函数作为网络的激活函数、均方误差（Mean Square Error，MSE）作为损失函数、自适应时刻估计算法（Adaptive Moment Estimation, Adam）作为优化器。

3.4 模型精度验证方法

本研究用于衡量模型性能的指标是可决系数（Coefficient of Determination，R^2）、均方根误差（Root Mean Square Error，RMSE）和平均绝对误差（Mean Absolute Error，MAE）。可决系数用于度量因变量的变异中可由自变量解释部分所占的比例，以此来判断统计模型的解释力；均方根误差可用来衡量模型估计值同观测值之间的偏差；平均绝对误差能更好地反映预测值误差的实际情况。具体计算公式为：

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (8)$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(f_i-y_i)^2}{n}} \quad (9)$$

$$MAE=\frac{\sum_{i=1}^{n}|f_i-y_i|}{n} \quad (10)$$

式中：y_i代表真实观测值；f_i代表模型预测值；$\bar{y}$代表观测值的平均值；y_i-f_i为残差，$\sum_{i=1}^{n}(y_i-f_i)^2$为残差平方和，$\sum_{i=1}^{n}(y_i-\bar{y})^2$为总平方和，$\sum_{i=1}^{n}(f_i-y_i)^2$为回归平方和。

本文将数据分成3份和7份，7份作为训练数据，3份作为测试数据。另外，为检测模型的过拟合情况，本文将数据集分成了10份，轮流将其中9份做训练1份做测试，10次训练的结果即为十倍交叉验证。本文计算了10倍交叉验证的可决系数（CV R^2）、均方根误差（CV *RMSE*）和平均绝对误差（CV *MAE*）。

普通克里格是气象数据空间化中常用的插值方法，为了进一步验证残差模型的优越性，本文还对3个气象因子分别做了普通克里格插值，并且计算了10倍交叉验证后的R^2、RMSE和MAE；为了验证模型结果的可靠性，还分别计算了残差模型结果和再分析资料同站点实测数据的皮尔逊相关系数。

4 结果及分析

4.1 数据统计

表2是3个气象数据的基本统计信息。表3是被解释变量与解释变量的皮尔逊相关性分析结果，日均气温同纬度的相关性最大；日均相对湿度与最近邻值的相关性最大，同纬度也有很大的相关性；日均风速同GLDAS风速的相关系数最大。

4.2 模型结果

表4是不同协变量组合下构建的模型的结果。气温模型列出了两组协变量，第2组在第1组的基础上加入了月份，无论是GAM方法还是残差网络方法，加入月份后模型的提升都不明显，研究选择第1组协变量参与模型构建；相对湿度模型有6组协变量，对GAM方法而言，除了最近邻值，其他的变量对模型的提升几乎没有帮助；对残差网络，气温和臭氧浓度可以小幅度改善模型，月份的加入有力的提升了模型的精度，研究选取第6组协变量参与模型构建；风速模型有3组协变量，GLDAS风速的加入大大提升了模型精度，而月份对模型几乎没有贡献，研究选择第2组协变量。

图2–图4分别为气温、相对湿度和风速的GAM和残差网络的训练结果和CV结果图，所有模型的CV结果和训练结果几乎没有差距，说明这些模型都没有过拟合。3个气象数据的残差网络模型都要好于GAM模型，且从散点图来看，残差网络模型估计的点相比于GAM模型更加集中在拟合线的附近。气温的GAM模型的CV R^2、RMSE、MAE分

表2 3个气象数据的基本信息

Tab. 2 Basic information of the three meteorological data

	记录数/个	最小值	最大值	平均值	中值	标准差
气温/℃	294 357	−37.70	38.20	12.97	14.90	11.45
高程/m		1.80	4612.20	770.00	361.90	953.04
相对湿度/%	290 925	4.00	100.00	67.21	71.00	19.57
研究所得气温/℃		−18.12	38.41	12.89	13.28	10.94
GEOS-FP臭氧浓度/DU		219.40	485.40	318.30	311.80	38.40
最近邻相对湿度/）		4.00	100.00	67.28	71.00	19.61
风速/（m/s）	255 209	0.00	23.20	2.06	1.80	1.27
GLDAS风速/（m/s）		0.32	19.22	2.80	2.40	1.58

表3 被解释变量和解释变量的皮尔逊相关系数

Tab. 3 Pearson's r between theexplained variables and explanatory variables

解释变量	日均气温	日均相对湿度	日均风速
经度	0.05	0.29	0.05
纬度	−0.45	−0.41	0.25
高程	−0.29	−0.35	0.13
DOY	0.13	0.20	−0.09
月份	−0.46	0.20	−0.13
研究所得气温		0.31	
GEOS-FP臭氧浓度		0.15	
最近邻值		0.89	
GLDAS风速			0.60

注：除特殊说明，表格中的相关系数都是在0.01水平下显著。

表4　气象数据各组协变量与模型结果

Tab. 4　Performance of each modelwithdifferent covariables

	协变量组合	GAM结果			残差自编码器结果		
		R^2	*RMSE*	*MAE*	R^2	*RMSE*	*MAE*
气温	经纬度+高程+DOY	0.87	4.05	3.10	0.95	2.47	1.87
	经纬度+高程+DOY+月份	0.87	4.06	3.10	0.96	2.26	1.71
相对湿度	经纬度+高程+DOY	0.51	13.77	10.96	0.72	10.37	8.05
	经纬度+高程+DOY+最近邻值	0.80	8.71	6.49	0.86	7.41	5.58
	经纬度+高程+DOY+气温	0.51	13.67	10.87	0.75	9.78	7.58
	经纬度+高程+DOY+臭氧浓度	0.52	13.59	10.79	0.75	9.74	7.55
	经纬度+高程+DOY+气温+臭氧浓度	0.52	13.64	10.81	0.77	9.47	7.29
	经纬度+高程+DOY+气温+臭氧浓度+月份	0.52	13.61	10.80	0.85	7.66	5.86
风速	经纬度+高程+DOY	0.22	11.27	7.81	0.44	9.55	6.60
	经纬度+高程+DOY+GEOS-FP风速	0.46	9.35	6.54	0.65	7.59	5.21
	经纬度+高程+DOY+GEOS-FP风速+月份	0.45	9.39	6.55	0.66	7.49	5.18

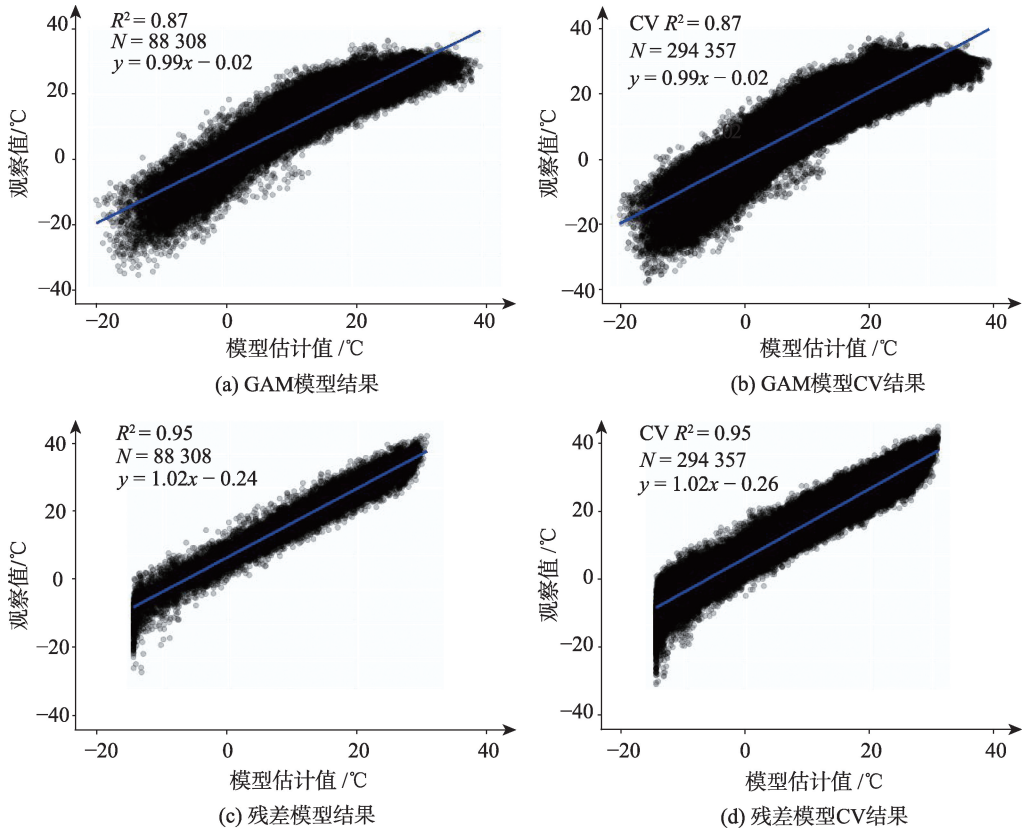


图2　2015年中国气温模型散点图

Fig. 2　Scatter plots of the temperature models in China in 2015

别为0.87、4.05 ℃、3.10 ℃，残差网络模型的CV R^2、*RMSE*、*MAE*分别为0.95、2.26 ℃、2.26 ℃；相对湿度的GAM模型的CV R^2、*RMSE*、*MAE*分别为0.52、13.59%、10.59%，残差网络模型的CV R^2、*RMSE*、*MAE*分别为0.85、7.53%、5.78%；风速的GAM模型的CV R^2、*RMSE*、*MAE*分别为0.45、0.84 m/s、0.65 m/s，残差网络模型的CV R^2、*RMSE*、*MAE*分别为0.66、0.74 m/s、0.51 m/s。

研究计算了3个气象因子站点观测数据同再分析资料和残差模型结果的皮尔逊相关系数：气温站点观测值同GLDAS气温和残差模型结果的相关系数分别为0.96和0.97，验证了残差模型生成的气温

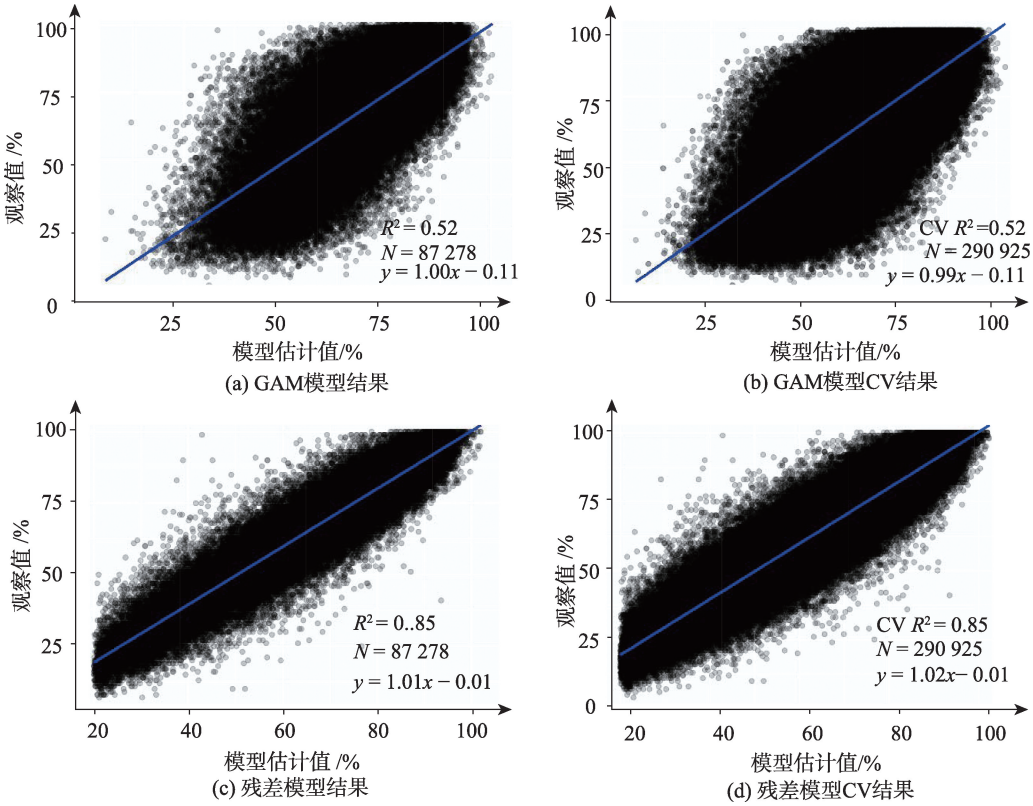


图3 2015年中国相对湿度模型散点图

Fig. 3 Scatter plots of the relative humidity models in China in 2015

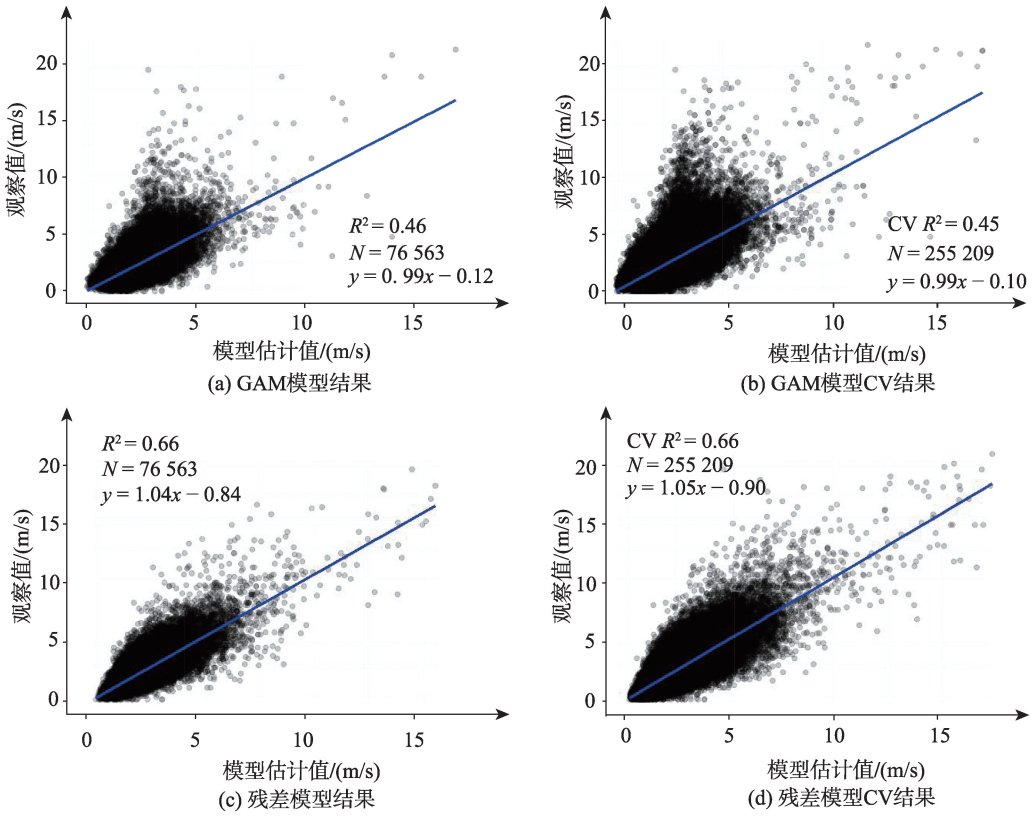


图4 2015年中国风速模型散点图

Fig. 4 Scatter plots of the wind speed models in China in 2015

数据的可靠性；风速站点观测值同GLDAS风速和残差模型结果的相关系数分别为0.49和0.70，残差模型生成的结果同风速站点观测数据具有更好的相关性；相对湿度站点观测值同NCEP/NCAR相对湿度和残差模型结果的相关系数分别为0.50和0.74，说明本研究的相对湿度模型结果更接近站点实测值。

从模型上而言，GAM模型是非线性回归，容易导致结果出现局部极值，而残差模型由于对称模型的应用，训练同测试的R^2及RMSE比较接近，不容易出现过拟合现象；从气象差异上而言，中国幅员辽阔，南北差异大，属于季风气候区，冬夏温差分布差异很大。冬季气温普遍偏低，南热北冷，南北温差大，超过50 ℃；夏季全国普遍高温（青藏高原除外），南北温差不大；中国从东南沿海到西北内陆，可划分为：湿润、半湿润、半干旱、干旱4类干湿地区；中国近地面风速在各区域差异很大。气候变化差异大的情况下采用残差模型更容易防止过拟合即极值的产生。本研究模型中的坐标及高程也可以很好地捕捉下垫面条件、地形条件，结果表明残差模型的精度目前是较好的。

4.3 重采样与格网效应

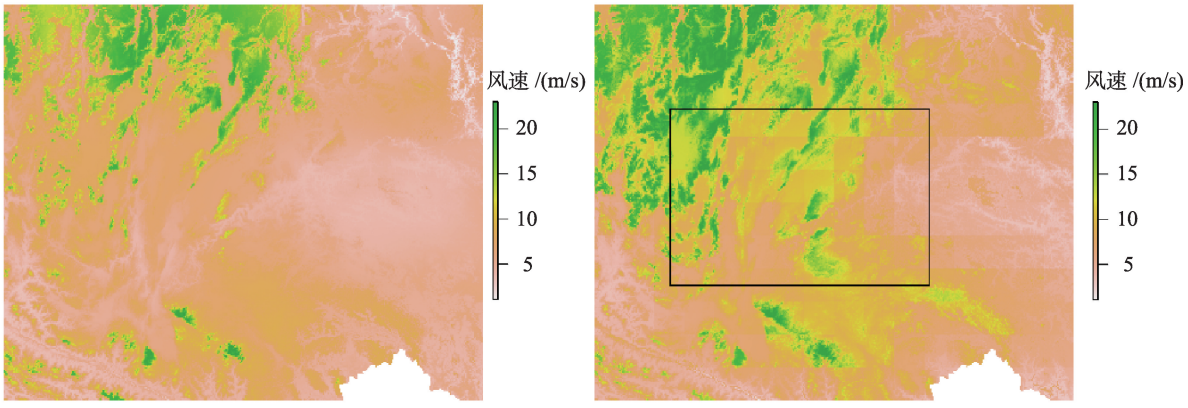
影像数据重采样前后，影像数据值与被解释变量的相关系数相差无几（如GLDAS风速重采样前相关系数为0.59，重采样后相关系数为0.60），且作为协变量在构建模型时各方面性能指标也几乎一样（R^2、RMSE、MAE）。图5展示了未重采样的GLDAS风速作为协变量和重采样后的GLDAS风速作为协变量分别参与模型的构建后，经模型估计得到的2015年1月1日的日均风速分布图。相比较于重采样后的结果，未重采样的结果图上覆盖有一层格网，值得注意的是，这层格网和GLDAS影像的格网重合，称之为格网效应。由于在风速的估计中用到了GLDAS影像数据，而这套数据在沿海及内陆都有些许的缺失值，因此研究生成的数据也在相同的位置存在缺失。

4.4 最近邻值与栅格多边形效应

最近邻值同相对湿度的相关性极大（皮尔逊相关系数接近0.9），同时，包含最近邻值的模型对被解释变量具有相当高的解释力（CV R^2达到0.85）。图6是使用残差网络模型结合相对湿度的第2组协变量（经纬度+高程+DOY+最近邻值）和第6组协变量（经纬度+高程+DOY+气温+臭氧浓度+月份）在2015年在1月1日生成的日均相对湿度分布图。比对没有使用最近邻值作为协变量的模型估计出来的日均相对湿度分布图，使用了最近邻值的分布图中有明显的多边形掩膜覆盖在栅格之上，称为多边形效应。最近邻值作为协变量是导致多边形效应的原因，若可以处理好最近邻值带来的多边形效应，那么最近邻值将会成为一个强有力的预测因子。

4.5 日均气象栅格图

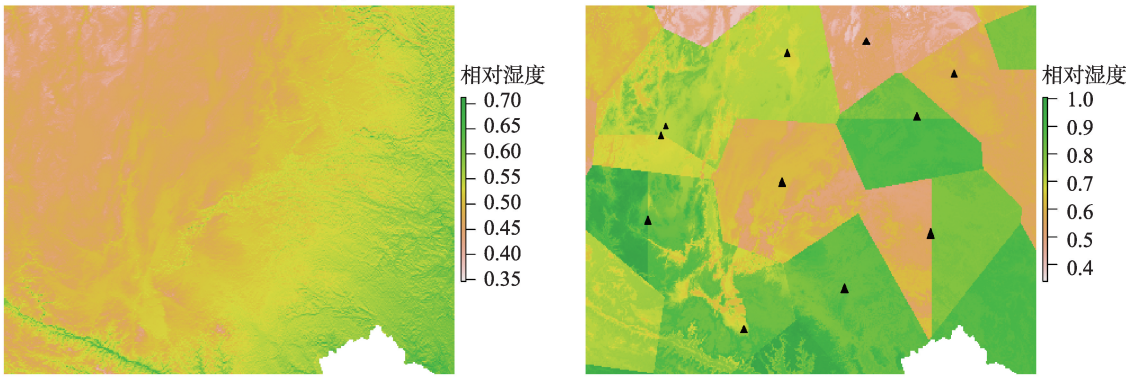
使用R软件生成2015年（365天）覆盖研究区域的1 km间隔的格网点数据，提取各模型所需的协变量，使用已经训练好的模型，估计出每一天各个气象数据的值。图7–图9分别为2015年1月1日的日均气温、日均相对湿度和日均风速分布图。在图7



(a) 未使用 GLDAS 影像　　(b) 使用 GLDAS 影像

图5　风速模型的格网化效应

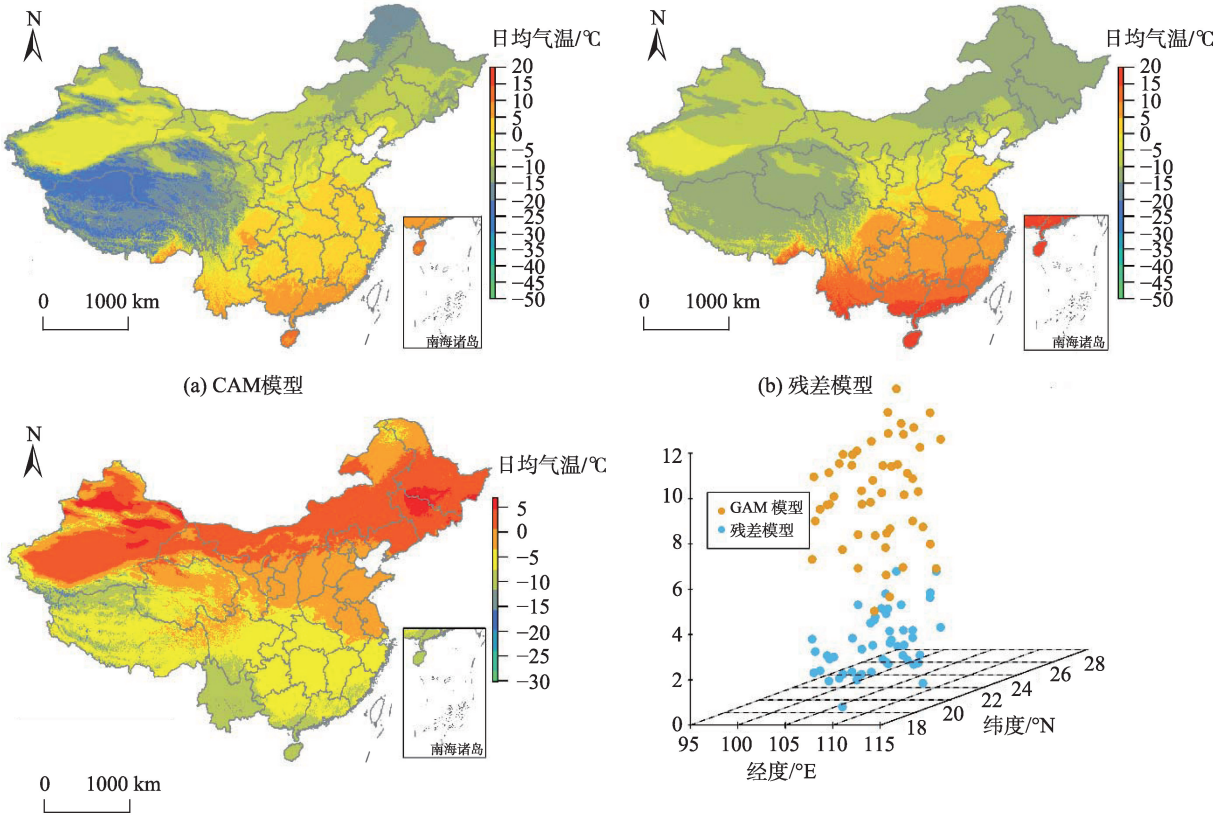
Fig. 5　Grid effects of the wind speed models



(a) 未使用近邻值　　(b) 使用近邻值

图6　相对湿度模型的多边形效应

Fig. 6　Polygon effects of the relative humidity models

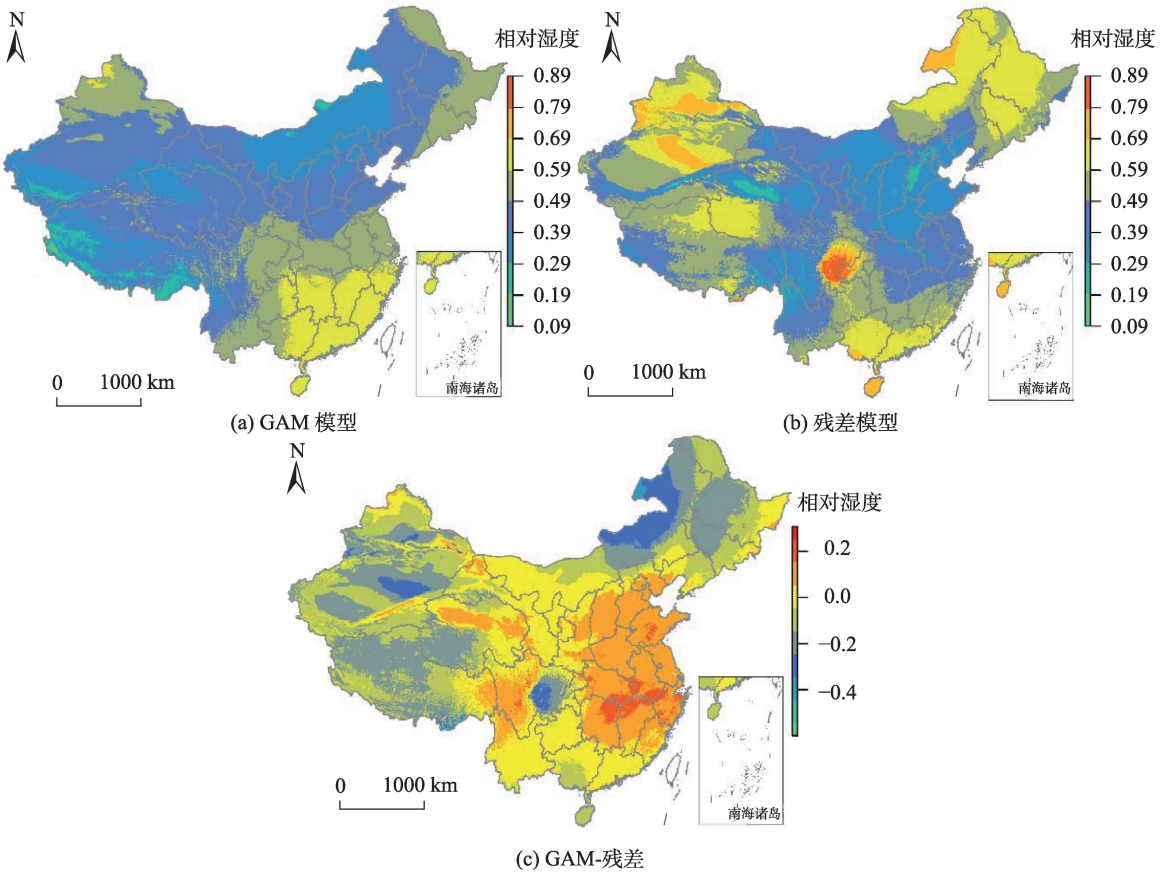


注：由于数据获取困难，本次研究不包括香港、台湾和澳门。该图基于国家测绘地理信息局标准地图服务网站下载的审图号为GS(2016)1570号的标准地图制作，底图无修改。

图7　2015年1月1日中国日均气温分布

Fig. 7　Daily average temperature on January 1, 2015

中残差网络模型的气温分布图中可以看到较为清晰的气温分界线，而GAM模型产生的分布图的分界线更加模糊，图8(c)是GAM模型结果和残差网络模型结果的差值，GAM模型的估计结果整体上低于残差网络模型，研究选取了两模型结果差距大于8 ℃的站点，做了图7(d)图：两模型差异大(相差超过8 ℃)的站点的残差图(模型结果和观测值的差值的绝对值)，GAM结果和观测值的差距普遍较



注：由于数据获取困难，本次研究不包括香港、台湾和澳门。该图基于国家测绘地理信息局标准地图服务网站下载的审图号为GS(2016)1570号的标准地图制作，底图无修改。

图8　2015年1月1日中国日均相对湿度分布

Fig. 8　Daily average relative humidity on January 1, 2015

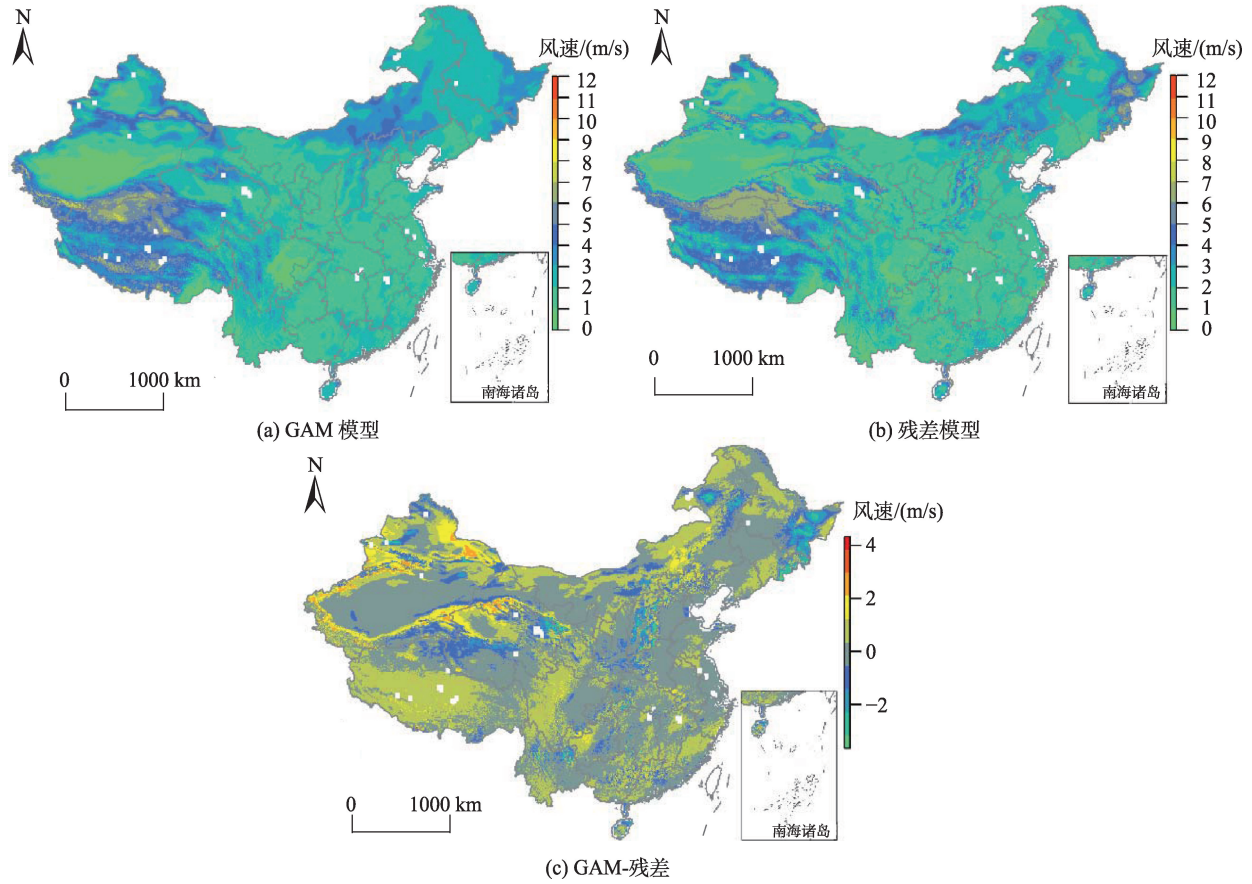
大，证明残差模型相较于GAM模型更接近真实值。GAM模型在气温值高的云南、海南等中国南部地区和气温值低的西部地区都存在估计值偏低的情况，即GAM方法在极值区域模型的估计能力差，估计值偏低；观察相对湿度分布图和风速分布图，残差网络模型结果显示的细节信息更加丰富。

4.6 普通克里格方法比对

为了同传统气象数据空间化方法进行比较，研究使用普通克里格（Ordinary Kriging，OK）对2015年1月1日的日均气温、日均相对湿度和日均风速进行了插值，10倍交叉验证的结果如表5所示，模型之间的差距随着变量估计难度的增大而增加，对于日均气温和日均相对湿度，普通克里格的结果比残差模型差，但是差距不那么明显，而对于风速而言，残差模型的表现力远好于普通克里格，模型的CV R^2增加了0.51，CV *RMSE*和CV *MAE*分别下降了36%和32%；除去精度上的优势外，残差模型只需训练一个模型，以天作为参数即可估计整整一年的气象数据，而普通克里格插值需要对每天的数据都构建一个模型，效率更低；此外，残差网络模型在进行格网估计时，速度远远快于普通克里格插值。

5　结论

本文采用了残差网络模型进行了高分辨率的气象因子（温度、相对湿度及风速）时空估计，获得了比经典的非线性回归模型GAM和传统插值模型普通克里格更好的结果，尤其在相对湿度及风速的预测方面。本文在模型中融入了时空因素，结果表明模型有效地捕捉了各要素的时空变化，取得了更好的估计精度。对于时间相关性的考虑，由于本文的地面样本为800多个，样本数量有限且每个样本时间相关性有差异，不适合采用类似长短期记忆（Long-Short-Term Memory，LSTM）时间序列网络



注:由于数据获取困难,本次研究不包括香港、台湾和澳门。该图基于国家测绘地理信息局标准地图服务网站下载的审图号为GS(2016)1570号的标准地图制作,底图无修改。

图9 2015年1月1日中国日均风速分布

Fig. 9 Daily average wind speed on January 1, 2015

表5 10倍交叉验证的OK插值结果

Tab. 5 Ten-fold cross-validation of the OK interpolation results

气象变量	模型	CV R^2	CV *RMSE*	CV *MAE*
日均气温	克里格	0.92	2.87	2.14
	残差模型	0.95	2.47	1.87
日均相对湿度	克里格	0.72	8.96	6.83
	残差模型	0.86	7.41	5.58
日均风速	克里格	0.14	11.78	7.67
	残差模型	0.65	7.59	5.21

深度学习模型。研究采用的残差模型,虽然直接将天作为时间要素捕捉时间相关性不是最好的方法,但也可以体现相邻时间样本点对目标点的影响,天类似于权重因子,最近(时间上)的点影响最大,在3个气象因子的预测中效果优于其他方法。此外,残差呈现随机性分布,说明模型捕捉到了主要的时间相关性。

气温是相对容易解释的变量,确定了地理位置、高程和DOY之后,使用简单的非线性回归就可以拟合出较好的模型。当使用残差网络方法时,模型的性能又有了进一步的提升,与GAM相比,RMSE和MAE都降低了将近一倍。

相对湿度的时空变异性更大,构建一个好的模型并且准确估计的难度更大。在模型本身的提升遇到瓶颈之后,加入一些相关变量来提升模型的表现力是常见的思路,日均气温和臭氧浓度的加入使得模型有了一定的改善,意外的是,月份因子的加入使得模型的R^2提高了相当多(0.77~0.85)。而这种现象在气温和风速模型中都没有发生,这暗示着合适的时间因子对模型的重要性。实际上,最近邻值可以为模型提供极大的贡献,但是最近邻值的加入同时也带来了结果栅格的多边形效应,在以后的研究中将试图在不引入多边形效应的情况下利用最近邻值。

风速是3个气象因子中最难估计的变量。GLDAS风速同站点风速有很高的相关性,参与模型构建可

以提升模型的性能。需要注意的是，使用影像之前需要对影像进行重采样处理，否则结果栅格会产生格网效应。GLDAS影像在中国的沿海区域和陆地有一些缺失值，这使得最后产生的数据也会存在缺失。今后本文希望可找到风速再分析资料的替代协变量，以期在保证模型精度的同时生成没有缺失的数据产品。另外，风速模型也考虑过最近邻值，但在加入最近邻值后模型精度基本没有提升。

被解释变量y由2个方面决定：解释变量x以及y与x的链接函数f。当总体存在空间分层异质性（Spatial Stratified Heterogeneity，SSH）并且f没有考虑SSH时，基于数据学习的估计将是混杂的。中国占地面积大，影响气象变量的因素复杂多样，为了进一步提升模型的性能，可以检验SSH后再确定如何使用模型，是全局使用还是分层使用。在未来的研究中可以利用王劲峰等[36-38]提出的地理探测器来检测总体是否存在SSH：如果SSH不显著，则全局统计模型是适用的；如果存在SSH，则可分别在每层使用模型，以避免估计混杂。

本研究成果为高分辨率的气象因素（气温、相对湿度及风速）的时空估计提供了重要的方法参考。而随着本文的方法的推广，产生覆盖全国的气象因素的栅格表面结果可为多种应用提供重要的数据来源。

参考文献(References)：

[1] Shi Y, Ho H C, Xu Y, et al. Improving satellite aerosol optical Depth-$PM_{2.5}$ correlations using land use regression with microscale geographic predictors in a high-density urban context[J]. Atmospheric Environment, 2018,190:23-34.

[2] Gao M, Yin L, Ning J. Artificial neural network model for ozone concentration estimation and Monte Carlo analysis [J]. Atmospheric Environment, 2018,184:129-139.

[3] Li L, Lurmann F, Habre R, et al. Constrained mixed-effect models with ensemble learning for prediction of nitrogen oxides concentrations at high spatiotemporal resolution[J]. Environmental Science & Technology, 2017,51(17):9920-9929.

[4] 王小燕,赵晓宇,陈恢富,等.江汉平原小麦孕穗期空气相对湿度升高的产量效应[J].中国农业科学,2014,47(19):3769-3779. [Wang X Y, Zhao X Y, Chen H F, et al. Characteristics of air moisture and the effects of high air moisture at booting stage on grain yield of wheat in Jianghan plain[J]. Scientia Agricultura Sinica, 2014,47(19):3769-3779.]

[5] 孙丽娜,孙瑞红,仇贵生,等.相对湿度对苹小卷叶蛾实验种群的影响[J].应用生态学报, 2014,25(12):3587-3592. [Sun L N, Sun R H, Qiu G S, et al. Influence of relative humidity on the Adoxophyes orana experimental population[J]. Chinese Journal of Applied Ecology, 2014,25(12):3587-3592.]

[6] 赵琴,潘静,曹兵,等.气温升高与干旱胁迫对宁夏枸杞光合作用的影响[J].生态学报,2015,35(18):6016-6022. [Zhao Q, Pan J, Cao B, et al. Effects of elevated temperature and drought stress on photosynthesis of Lycium barbarum[J]. Acta Ecologica Sinica, 2015,35(18):6016-6022.]

[7] Bo Y C, Song C, Wang J, et al. Using an autologistic regression model to identify spatial risk factors and spatial risk patterns of hand, foot and mouth disease (HFMD) in Mainland China[J]. BMC Public Health, 2014,14(1):358.

[8] 谢今范,刘玉英,王玉昆,等.东北地区风能资源空间分布特征与模拟[J].地理科学,2014,34(12):1497-1503. [Xie J F, Liu Y Y, Wang Y K, et al. Spatial distribution characteristics of wind resource and its simulation in northeast China[J]. Scientia Geographica Scinica, 2014,34(12):1497-1503.]

[9] 任永建,刘敏,袁业畅,等.湖北省风能资源的高分辨率数值模拟试验[J].自然资源学报,2012,27(6):1035-1043. [Ren Y J, Liu M, Yuan Y C, et al. The high resolution numerical simulation of wind energy resource in Hubei Province[J]. Journal of Natural Resources, 2012,26(6):1035-1043.]

[10] 于洋,卫伟,陈利顶,等.黄土高原年均降水量空间插值及其方法比较[J].应用生态学报,2015,26(4):999-1006. [Yu Y, Wei W, Chen L D, et al. Comparison on the methods for spatial interpolation of the annual average precipitation in the Loess Plateau region[J]. Chinese Journal of Applied Ecology, 2015,26(4):999-1006.]

[11] 王国泰,张守平,杨清伟,等.基于空间插值方法的重庆降水信息展布[J].南水北调与水利科技,2018,16(3):18-23. [Wang G T, Zhang S P, Yang Q W et al. Precipitation information distribution in Chongqing based on spatial interpolation method[J]. South-to-North Water Transfers and Water Science & Technology, 2018,16(3):18-23.]

[12] 林忠辉,莫兴国,李宏轩,等.中国陆地区域气象要素的空间插值[J].地理学报,2002,57(1):47-56. [Lin Z H, Mo X G, Li H X, et al. Comparison of three spatial interpolation methods for climate variables in China[J]. Acta Geographica Sinica, 2002,57(1):47-56.]

[13] 张仁平,张云玲,郭靖,等.新疆地区降水分布的空间插值方法比较[J].草业科学,2018,35(3):521-529. [Zhang R P, Zhang Y L, Guo J, et al. Comparison of spatial interpolation methods for precipitation distribution in Xinjiang region[J]. Pratacultural Science, 2018,35(3):521-529.]

[14] 姜晓剑,刘小军,黄芬,等.逐日气象要素空间插值方法的比较[J].应用生态学报,2010,21(3):624-630. [Jiang X J, Liu X J, Huang F, et al. Comparison of spatial interpolation methods for daily meteorological elements[J]. China Journal of Applied Ecology, 2010,21(3):624-630.]

[15] 陈思宁,郭军.不同空间插值方法在区域气温序列中的应用评估:以东北地区为例[J].中国农业气象,2015,36(2):234-241. [Chen S N, Guo J. Evaluation of different spatial interpolation methods in regional temperature sequence: A case study in northeast China[J]. China Journal of Agrometeorology, 2015,36(2):234-241.]

[16] 彭彬,周艳莲,高苹,等.气温插值中不同空间插值方法的适用性分析——以江苏省为例[J].地球信息科学学报,2011,13(4):539-548. [Peng B, Zhou Y L, Gao P, et al. Suitability assessment of different interpolation methods in the gridding process of station collected air temperature: A case study in Jiangsu Province, China[J]. Journal of Geo-information Science, 2011,13(4):539-548.]

[17] 李军龙,张剑,张丛,等.气象要素空间插值方法的比较分析[J].草业科学,2006(8):6-11. [Li J L, Zhang J, Zhang C, et al. Analyze and compare the spatial interpolation methods for climate factor[J]. Pratacultural Science, 2006(8):6-11.]

[18] 李伟,李庆祥,江志红.用Kriging方法对中国历史气温数据插值可行性讨论[J].南京气象学院学报,2007(2):246-252. [Li W, Li Q X, Jiang Z H. Discussion on feasibility of gridding the historic temperature data in China with kriging method[J]. Journal of Nanjing Institute of Meteorology, 2007(2):246-252.]

[19] Xu C D, Wang J F, Li Q. A new method for temperature spatial interpolation based on sparse historical stations[J]. Journal of Climate, 2017,31(5):1757-1770.

[20] Wang J F, Xu C, Hu M, et al. Global land surface air temperature dynamics since 1880[J]. International Journal of Climatology, 2018,381:E466-E474.

[21] Wang J F, Xu C D, Hu M G, et al. A new estimate of the China temperature anomaly series and uncertainty assessment in 1900-2006[J]. Journal of Geophysical Research: Atmospheres, 2014,119(1):1-9.

[22] 解恒燕,张深远,侯善策,等.降水量空间插值方法在小样本区域的比较研究[J].水土保持研究,2018,25(3):117-121. [Xie H Y, Zhang S Y, Hou S C, et al. Comparison research on rainfall interpolation methods for small sample areas[J]. Research of Soil and Water Conservation, 2018, 25(3):1-9.]

[23] 马秀霞,黄领梅,沈冰.陕西省月平均气温空间插值方法研究[J].水资源与水工程学报,2017,28(5):100-105. [Ma X X, Huang L M, Shen B Study on spatial interpolation method of monthly mean temperature in Shanxi Province [J]. Journal of Water Resource & Water Engineering, 2017,28(5):100-105.]

[24] 沈红,刘文兆,张勋昌.黄土高原气象要素栅格化方法的研究[J].西北农林科技大学学报(自然科学版),2010,38(4):99-106. [Shen H, Liu W Z, Zhang X C. Studying the methods for rasterizing meteorological variables in the Loess Plateau[J]. Journal of Northwest A&F University (Nat. Sci. Ed.), 2010,38(4):99-106.]

[25] 李月臣,何志明,刘春霞.基于站点观测数据的气温空间化方法评述[J].地理科学进展,2014,33(8):1019-1028. [Li Y C, He Z M, Liu C X. Review on spatial interpolation methods of temperature data from meteorological stations[J]. Progress in Geography, 2014,33(8):1019-1028.]

[26] 周婷婷,陈文惠.基于MODIS数据和气象观测数据的气温空间插值方法比较[J].地理科学进展,2011,30(9):1143-1151. [Zhou T T, Chen W H. Comparation of the temperature spatial interpolation methods based on MODIS data and meteorological observation data[J]. Progress in Geography, 2011,30(9):1143-1151.]

[27] 石志华,刘梦云,常庆瑞,等.基于优化参数的陕西省气温、降水栅格化方法分析[J].自然资源学报,2015,30(7):1141-1152. [Shi Z H, Liu M Y, Chang Q R, et al. Comparison of temperature and precipitation rasterization methods based on optimized parameters in Shanxi Province[J]. Journal of Natural Resources, 2015,30(7):1141-1152.]

[28] 姜燕敏,吴昊旻. 20个CMIP5模式对中亚地区年平均气温模拟能力评估[J].气候变化研究进展,2013,9(2):110-116. [Jiang Y M, Wu H M. Simulation capabilities of 20 CMIP5 methods for annual mean air temperatures in central Asia[J]. Progress Inquisitions de Mutatione Climatis, 2013,9(2):110-116.]

[29] Lazzus J A. Estimation of surface soil temperature based on neural network modeling[J]. Italian Journal of Agrometeorology-rivista Italiana Di Agrometeorologia, 2014, 19(2):5-12.

[30] Szymanowski M, Kryza M, Spallek W. Regression-based air temperature spatial prediction models: An example

from Poland[J]. Meteorologische Zeitschrift, 2013,22(5): 577-585.

[31] Hjort J, Suomi J, Kayhko J. Spatial prediction of urban-rural temperatures using statistical methods[J]. Theoretical and Applied Climatology, 2011,106(1-2):139-152.

[32] Kou X, Jiang L, Bo Y, et al. Estimation of land surface temperature through blending MODIS and AMSR-E data with the Bayesian Maximum Entropy method[J]. Remote Sensing, 2016,8(2):105.

[33] 蒋冲,王飞,刘焱序,等.秦岭南北风速时空变化及突变特征分析[J].地理科学,2013,33(2):244-250. [Jiang C, Wang F, Liu Y X, et al. Spatial-temporal variation and mutation of wind speed in the northern and southern regions of the Qinling Mountains[J]. Scientia Geographica Sinica, 2013,33(2):244-250.]

[34] 赵伟,高博,刘明,等.气象因素对香港地区臭氧污染的影响[J].环境科学,2019,40(1):55-66. [Zhao W, Gao B, L Q, et al. Impact of meteorological factors on the ozone pollution in Hong Kong[J]. Environmental Science, 2019,40 (1):55-66.]

[35] Li L, Y F, J W, et al. Autoencoder based residual deep networks for robust regression prediction and spatiotemporal estimation[J]. 2018, preprint arXiv:1812.11262.

[36] Wang J F, Zhang T, Fu B. A measure of spatial stratified heterogeneity[J]. Ecological Indicators, 2016,67:250-256.

[37] 王劲峰,徐成东.地理探测器:原理与展望[J].地理学报, 2017,72(1):116-134. [Wang J F, Xu C D. Geodector: principle and prospective[J]. Acta Geographica Sinica, 2017,72(1):116-134.]

[38] Wang J F, Li X H, Christakos G, et al. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China [J]. International Journal of Geographical Information Science, 2010,24(1):107-127.