利用集合深度学习方法融合多源数据开发全国能见度网格数据

吕宝磊[1]　胡泳涛[2]　李林[3]　梁海河[4]　刘钧[1]　王晓江[5]

（1 华云升达（北京）气象科技有限责任公司，北京 102299；2 佐治亚理工学院土木与环境工程学院，亚特兰大 30332；3 北京市气象局，北京 100089；4 中国气象局气象探测中心，北京 100081；5 中国华云气象科技集团公司，北京 100081）

摘要：能见度对于生产生活安排具有重要的指导意义，目前我国的能见度监测网络站点覆盖相对稀疏，观测数据具有空间离散性和局地性。利用集合深度学习模型融合多源数据估计能见度，该集合模型包括了深度神经网络模型、随机森林模型、梯度强化模型和广义线性模型四种机器学习器，融合了气象观测、大气成分观测、模式模拟和土地利用类型等多源数据，并利用Barnes客观分析进一步消除深度学习融合误差，开发出空间连续的能见度网格数据集。该数据的空间分辨率为12 km，经过独立样本评估，融合数据的R^2为0.61。相比较于空间插值和线性模型等方法，提出的方法具有更好的准确度，且具有很好的空间解析力，可进一步用于开发更高分辨率数据。该方法具有较高的计算效率和较好的数据兼容性，可以部署于业务化平台中可靠运行。

关键词：深度学习，能见度，多源数据融合，Barnes客观分析

DOI：10.3969/j.issn.2095-1973.2018.06.008

Fuse Multiple Data Sources with an Ensemble Deep Learning Approach to Estimate Nationwide Gridded Visibility

Lü Baolei[1], Hu Yongtao[2], Li Lin[3], Liang Haihe[4], Liu Jun[1],Wang Xiaojiang[5]

(1 Huayun Sounding Meteorological Technology Company, Ltd., Beijing 102299

2 School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

3 Beijing Meteorological Service, Beijing 100089

4 Meteorological Observation Center, China Meteorological Administration, Beijing 100081

5 China Huayun Meteorological Technology Group, Beijing 100081)

Abstract: Ground visibility had significant influences on transportation and human outdoor activities. Stationary visibility observations are usually spatially sparse, inadequate to meet the demand for complete spatial coverage data. To tackle the problem, we put forward a novel data fusion framework by incorporating an ensemble deep learning method and a Barnes objective analysis approach. The ensemble deep learning method works to generate high-quality first-guess field by building complex linkages between data sets. The Barnes objective analysis method was used to remove remaining residual errors and biases. Furthermore, it was designed to be hierarchical to better address the environmental factors such as particulate concentrations and relative humidity that greatly affect visibility in a sub-layer. The data fusion framework is able to include multiple data sources of different types, i.e. gridded WRF meteorological predictions, CMAQ air quality predictions, stationary meteorological and atmospheric composition observations and other supporting land use/cover data sets. The model was implemented to generate gridded hourly visibility data sets in China. The fused gridded data was evaluated against observations from independent monitors, with an R^2 = 0.61. Comparatively the R2 values of interpolation approach and linear regression fusion were respectively 0.55 and 0.57. Besides, our gridded fusion data have much more and detailed spatial information than that of the smoothed interpolation data. Our data fusion approach is relatively easy to implement in operational system and also has good extensibility to handily include more data sets.

Keywords: deep learning, visibility, data fusion, Barnes objective analysis

收稿日期：2018 年 6 月 1 日；修回日期：2018 年 10 月 19 日

第一作者：吕宝磊（1988—），Email：baoleilv@foxmail.com

0　引言

近地面能见度对于人们的生产生活、交通物流具有重要的影响。由于工业化和城市化进程的加快，我

国的近地面水平能见度自20世纪90年代以来呈现持续下降的趋势[1−2]，大气通透度同时也在降低[3]。同时以颗粒物污染为特征的雾霾事件频繁发生，雾霾过程中能见度甚至不足百米，严重影响了正常的出行活动。

能见度变化与大气成分对可见光的消光作用相关，消光作用包括散射作用和吸收作用[4]。大气中水汽和气溶胶颗粒物成分是影响能见度最重要的两个因素[5]。大气中细颗粒物（$PM_{2.5}$）浓度与能见度呈现负相关的关系，同时这种关系受到湿度负相关调节[6−8]。近年来随着$PM_{2.5}$为代表的颗粒物污染的增加[9−10]，气溶胶成分对能见度下降起到越来越重要的作用。$PM_{2.5}$复杂的化学组分使得其具有显著的消光作用[6]，比如无机盐组分对可见光的散射作用，黑炭和其他含碳组分的光吸收作用。在湿度较大的情况下，气溶胶组分会发生吸湿性增长，使得其消光作用成倍增长[11]。白永清等[10]利用幂函数在武汉拟合了不同的湿度下的相关关系，相关系数达到0.8，拟合得到的幂函数的指数在−0.75～−0.6。樊高峰等[12]在杭州也利用幂函数拟合了$PM_{2.5}$和能见度之间的关系，得到的R^2在0.7左右，幂函数指数在−0.8左右。王淑英等[13]利用对数函数在北京拟合了可吸入颗粒物（PM_{10}）和能见度之间的关系，得到的R^2在0.7左右。由此可见，$PM_{2.5}$浓度和相对湿度是影响能见度的最主要因素。

我国已经建立起来了包括2000余个站点的能见度监测网络，相比较于我国广袤的国土面积，这些监测站点仍旧缺乏足够的空间代表性。同时，对能见度影响较大的$PM_{2.5}$浓度和相对湿度分布具有显著的空间异质性[14]，进一步降低了能见度监测的代表性。因此，估计出缺乏监测站分布地区的能见度，开发一种空间连续且具有较高分辨率以及较好准确性的网格化能见度数据，将具有十分重要的应用价值。

估计能见度的方法主要有两种：一种是分别考虑气溶胶的每种组分在一定湿度下的消光作用，例如利用IMPROVE公式计算[15−16]；另外一种是将各种因素，尤其是$PM_{2.5}$浓度和相对湿度，放入统计模型中对能见度进行模拟[17−18]。目前大多数的研究是在单一点位和城市进行能见度模拟和预测分析[7, 10]，且主要集中在时间上的预报。例如白永清等[10]利用神经网络算法，输入$PM_{2.5}$和相对湿度数据，对点位上的逐小时能见度进行模拟。其训练相关系数R为0.92（R^2 = 0.82），预报时R达到了0.86（R^2=0.74），证明了神经网络在能见度拟合方面的有效性。Zhu等[19]利用深度神经网络模型在乌鲁木齐机场开展了能见度与预报研究，得到了能见度的趋势预报。全国范围内的能见度模拟，尤其是空间上的模拟研究依然较少。有限的深度学习方法的应用案例也证明了其在能见度模拟预报方面具有优势。本文利用集合深度学习和残差插值的方法融合了多种模式模拟数据、观测数据和土地利用数据，开发出了12 km分辨率的全国能见度逐小时数据。该方法具有较好的准确性，该数据目前已经准业务化生产，可以为交通出行等领域提供可靠的决策支撑。

1 资料与方法

1.1 观测与模拟数据

1）气象观测数据

逐小时能见度观测数据来源于国家基本气象站和有能见度观测的交通站。国家基本气象站分布在全国东部地区，且空间分布较为均匀，共有2800多个。逐小时能见度包括1 min能见度和10 min能见度两种，使用10 min能见度以确保更广的时间覆盖。

湿度数据来源于国家基本站和区域站，一共有5万多个区域自动站的数据，经过质控之后的有效数据在每小时3万多条数据。

2）$PM_{2.5}$观测数据

$PM_{2.5}$观测数据有两个来源：一个是中国环境监测中心建立的空气质量监测网络，包括1493个监测站；另外一个是气象局建立的大气成分监测站，包括263个气象局大气成分监测站。这两个监测网络均可以提供逐小时的业务化的$PM_{2.5}$监测数据。在全国中东部和东北地区有较为均为的分布。两个网络一共有近1800个监测站点可使用，但这些监测站主要分布在城市地区，其空间代表性不如气象监测站。

3）气象模拟数据

气象模拟数据来自于WRF模型，利用GFS预报数据进行驱动。模拟的网格分辨率为12 km，投影方式为兰伯特等角投影。本文使用的气象数据为近地面的温度、湿度、风速以及边界层高度数据。

4）空气质量模拟数据

空气质量模拟数据来源于WRF−CMAQ模型，这里CMAQ模型版本为v5.0.1[20]，排放清单是通过高阶敏感性分析工具分析制作的动态清单。空气质量模型的模拟网格设置与气象模拟的设置一样。本文使用的是业务化的预报模型系统，其可以逐日做出未来120 h的空气质量预报。为了确保空气质量模拟的准确性，本文使用前24 h的空气质量预报结果。

除了上述数据之外，研究还使用了数字高程数据，城市覆盖度数据和林地覆盖度数据。

1.2 数据融合方法

在开展数据融合之前，对影响能见度的变量进行了统计分析，识别出对能见度影响较大的变量要素，

本文发现能见度与湿度和颗粒物浓度相关性较强，这与前人的大量分析一致[21]。为了更好地将能见度的影响因素考虑到模型中来，本文使用了两层融合的方法（图1）。在第一层融合模型中，首先将对能见度影响较大的$PM_{2.5}$浓度和相对湿度进行了融合，在同一网格设置下开发出具有较高准确度的能见度数据产品。在$PM_{2.5}$融合方法上，使用了集合深度学习和残差空间插值的方法，基于$PM_{2.5}$浓度观测值与模拟值及其他气象和土地利用数据，得到融合数据产品，该产品的准确度R^2在0.7左右[22]。对于湿度数据的融合，由于湿度监测点位在空间上密度较大，在WRF模拟相对湿度和观测值的基础上，直接采用了最优插值方法[23]，获得了相对湿度的融合分析场。

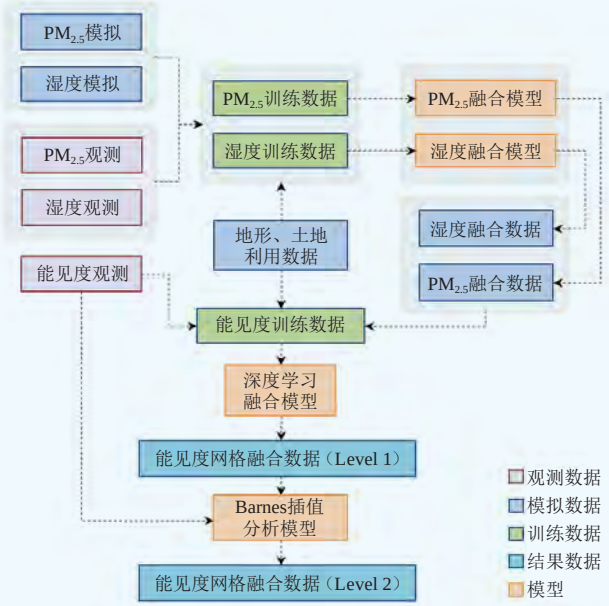

图1 网格化能见度实况分析数据开发方法

Fig. 1 The approach to develop gridded visibility fusion data

在第二层融合模型中，通过集合深度学习方法模拟能见度。集合深度学习方法综合了深度神经网络（Deep neural network，DNN）、随机森林模型（Random forest，RF）、广义线性模型（General linear model，GLM）和梯度提升模型（Gradient boosting machine，GBM）四个模型作为主学习器（图2）。在四个模型中，深度神经网络模型具有复杂非线性关系的拟合能力，其统计回归的结果具有无偏性，但存在过拟合的可能性[24]。随机森林模型和梯度提升模型本身均是包含了多种弱分类器的集合模型，只是随机森林模型使用了bootstrap aggregating （bagging）的方法[25]，梯度提升模型使用了Boosting的方法[26]，这两种方法在选择合适数量的决策树的情况下均可避免明显的过拟合现象。广义线性模型是线性模型的扩展，通过概率分布函数来实现对非线性过程的模拟，该模型不会出现明显的过拟合，效果稳定，但模拟的误差一般较大。具体来讲，深度神经网络模型设计为一个三层的全连接神经网络模型，以双曲正切函数为激活函数、Sigmoid函数为输出层函数；随机森林模型和梯度提升模型分别包含100颗分类树，激活函数也为双曲正切函数；广义线性模型使用的联系函数为高斯函数。

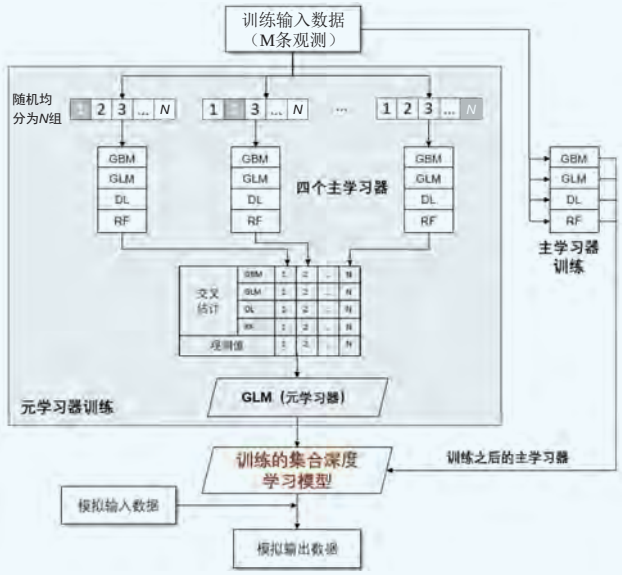

图2 集合深度学习模型训练与模拟过程

Fig. 2 Fitting and prediction process by ensemble deep learning approach

为了将四个学习器整合到一起，引入了一个元学习器，本文采用了GLM为元学习器。使用GLM作为元学习器的原因是它具有清晰的变量权重系数结构，可以对各个主学习器的效应产生清晰的认识，同时能够模拟非线性的响应关系。元学习器和主学习器的训练方法不同且较为复杂，它首先将M条训练数据随机采样分为N个批次，然后循环将其中的$N-1$个批次输入到模型中，并对剩余的1个批次做模拟。如此循环N次，即可得到每个主学习器在所有M条训练数据所对应的模拟值，然后将4列模拟值与观测值输入到元学习器中，训练得到每个模型的权重系数和偏差[27]。本文中采用的循环次数为10。集合深度学习比单独的深度学习模型的表现更好，这在其他研究[27]中也已经得到了证明。

通过该集合学习模型，以能见度为被解释变量，以融合的湿度、$PM_{2.5}$浓度和其他地形、土地利用、气象数据为基础，完成对能见度的模拟。这时获得的能见度还存在一定的模拟误差，然后将站点的训练误

差通过Barnes客观分析法对能见度模拟值做进一步修正[28−29]，提高能见度准确度，并提升网格化能见度的空间解析度，作为最终的能见度数据产品。目前开发的能见度数据的空间分辨率为12 km，未来该模型框架将在4 km和1 km的空间分辨率的网格上进行开发。

2 结果与讨论

2.1 能见度与湿度、$PM_{2.5}$浓度关系探讨

通过对北京市南郊站点的相对湿度、$PM_{2.5}$浓度和能见度之间的关系进行分析（图3），可以发现，能见度和$PM_{2.5}$浓度存在明显的负相关关系，在不同的相对湿度范围这种关系存在着差异，在不同的湿度下利用幂函数拟合了相关关系，得到的幂函数的指数在−1左右，拟合的R^2在0.7左右。由此可见，在相对湿度相对稳定的情况下，颗粒物对能见度变化的贡献能够达到70%～80%，这与以往在杭州、北京和武汉地区的研究结果也较为一致[7−8, 10]。因此在估计能见度时，获得可靠的$PM_{2.5}$浓度和相对湿度变得十分关键，这也是本文中使用$PM_{2.5}$浓度和相对湿度数据融合子模型的原因。

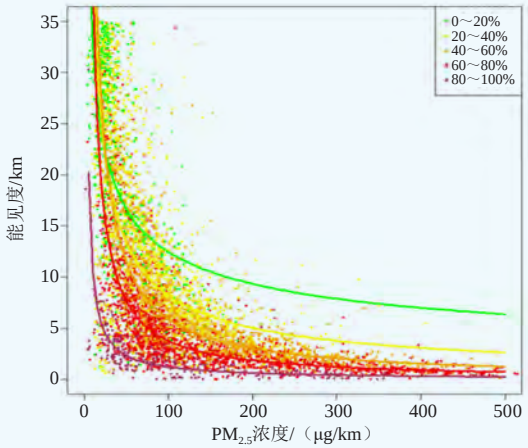

图3 2016年5月—2017年4月北京气象站点逐小时$PM_{2.5}$浓度、相对湿度和能见度之间的关系

Fig. 3 The relationship between $PM_{2.5}$ concentrations and visibility with the relative humidity in hourly data at a meteorological observation station in Beijing for May 2016–April 2017

2.2 能见度模拟与评估

利用2016年1月的逐小时数据进行了测试评估，每次随机采样10%的站点数据作为测试数据集，剩余90%数据进行能见度产品开发，得到的结果与10%的测试数据进行比对。对比了三种方法，即线性回归加Barnes客观分析、集合深度学习加Barnes客观分析和只进行克里金空间插值。由图4可以看出，三种方法的结果呈现一致的变化趋势。当能见度模拟效果较好时，利用集合深度学习的方法的效果最好，空间插值的效果最差。随着模拟表现的下降，三种方法的表现趋于一致。综合来看，在该时间范围内，集合模型的效果最好，平均R^2可以达到0.61，比插值模型的0.55高出11%，比使用线性模型的结果0.57高出6%。

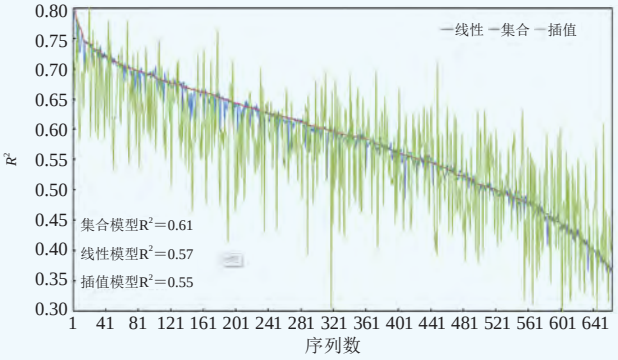

图4 不同方法获得的能见度模拟效果独立评估

Fig. 4 Evaluation of the accuracy of estimated visibility by different methods

如引言部分所讨论的，之前很多研究在单一站点的预报准确度R^2能够达到0.7甚至0.8。但仔细分析，他们研究结果的表现在业务化系统中并不能实现，首先这是因为他们使用的是观测相对湿度和观测$PM_{2.5}$做的测试，实际业务环境中我们只能使用相对湿度和$PM_{2.5}$浓度的预报值进行预测，真实表现应有所下降。另外，本文中的能见度空间模拟还要求模型具有空间扩展性，为了反映模型在真实应用环境下的表现，严格确保方法开发和测试环境和业务化应用环境一致，如气象和空气质量模拟场为业务化预报场，逐小时实时观测训练模型而没有用到时间超前数据做训练等，因此本文中的评估测试结果具有较好的可靠性，将与模型在业务化系统中的表现基本一致。

利用多源数据融合的多步骤的方法可以提高模拟的精度，另外该方法对能见度空间特征细节方面有更好的解析（图5）。通过基于深度学习的实时能见度产品可以更好的模拟能见度的空间分布特征。比如在关中地区，通过能见度插值的空间分布更加弥散，不能够反映出大量的人为活动只聚集在谷底地区而造成的局地性能见度降低，而通过深度学习方法获得的能见度变化更加清晰，这与利用卫星反演出来的$PM_{2.5}$浓度空间分布特征更加吻合[30]。另外，在河北省南部地区，沿着太行山的平原地区是污染排放和积累都非常严重的地区，细颗粒物浓度常年很高，然而该地区西部山区的能见度又较高，因此该地区能见度变化较为剧烈。利用集合深度学习的方法可以更好地表征山脉和平原交界处的走向，而直接插值的空间变化呈现梯

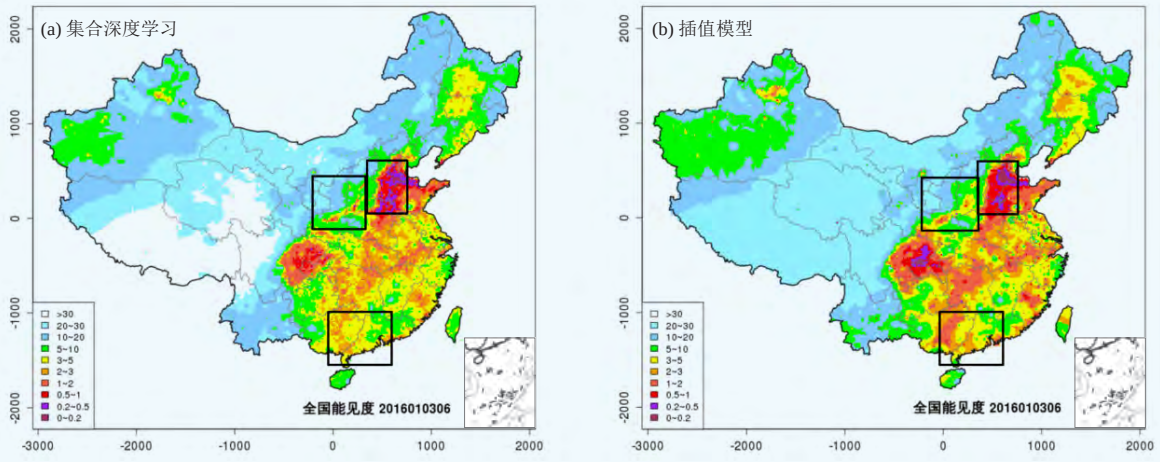

图5 同一时间（2016年1月3日06时）低能见度场景下融合模拟和插值效果对比

Fig. 5 Comparison of the gridded visibility data beween the fusion and interpolation methods in a typical low visibility scenario (at 06 BT on January 3 2016)

度特征，具有较差的空间解析度，深度学习数据融合方法可以更好地反映能见度的变化特征。

通过多源数据融合方法，可以构建出一个实时的能见度融合数据集，形成网格化能见度数据序列。图6展示了连续8 h的能见度空间分布特征，由此可以看出该方法得到的结果在时间上有较好的连续性和稳定性，确保其在实际应用中具有良好的可靠性。基于本方法利用R语言相关算法库和Shell脚本，已经实现了业务化稳定运行。本文的方法测试中只使用了90%的站点数据，在业务化系统中使用全部的数据之后，模型的模拟效果将有进一步的提升。

3 小结

本文提出了一种基于深度学习和Barnes客观分析法的多源数据融合方案来融合多源观测数据，该方案可以融合多种类型的模式数据、站点观测数据和土地利用等其他数据。

通过2016年1月逐小时数据的测试，利用深度学习方法的数据融合具有更好的准确性，得到的R^2为0.61，明显优于插值得到的结果R^2=0.55。同时模型结果具有更好的空间解析度。测试结果可以反映模型业务化实践的准确性。

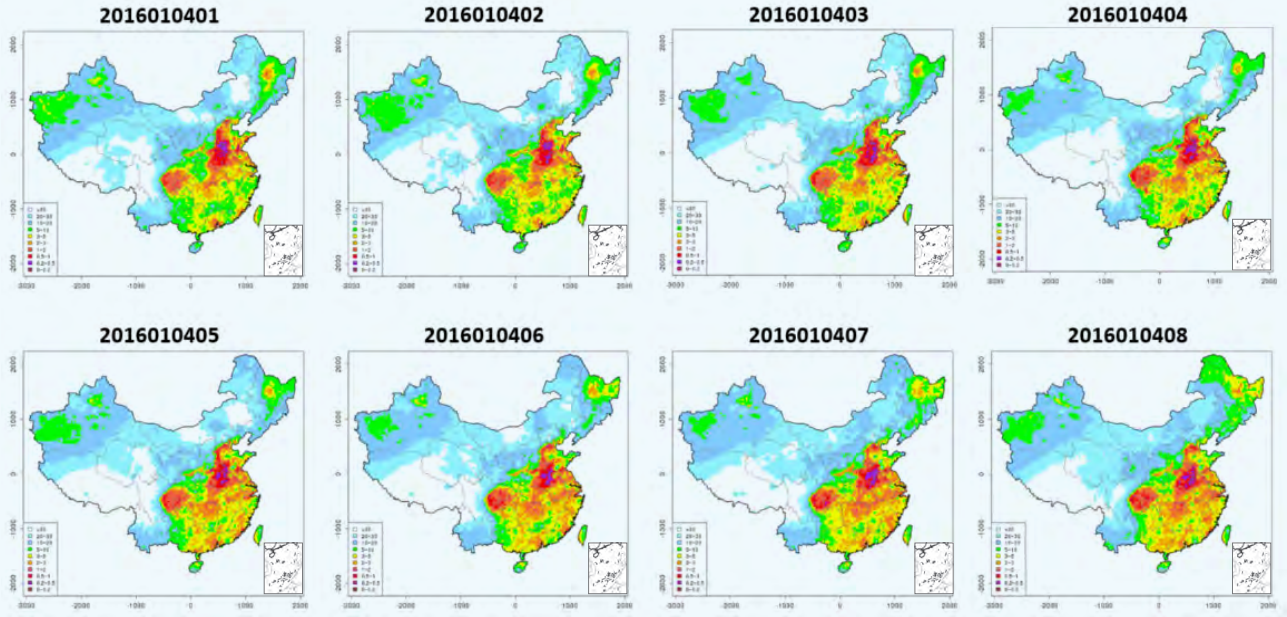

图6 2016年1月4日01—08时全国能见度模拟结果时间序列示例

Fig. 6 The spatio-temporal time-series (at 01–08 BT on January 3, 2016) of the fused gridded visibility in China

参考文献

[1] Che H, Zhang X, Li Y, et al. Horizontal visibility trends in China 1981-2005. Geophysical Research Letters, 2007, 34(24): 497-507.
[2] Chen H, Wang H. Haze days in North China and the associated atmospheric circulations based on daily visibility data from 1960 to 2012. Journal of Geophysical Research: Atmospheres, 2015, 120(12): 5895-5909.
[3] Wang K, Dickinson RE, Liang S. Clear sky visibility has decreased over land globally from 1973 to 2007. Science, 2009, 323(5920): 1468.
[4] Chow J, Bachmann J, Wierman SG, et al. Visibility: science and regulation. Air Repair, 2002, 52(9): 973-999.
[5] 陈静, 赵春生. 大气低能见度的影响因子分析及计算方法综述. 气象科技进展, 2014(4): 44-51.
[6] Chen J. Impact of relative humidity and water soluble constituents of PM2.5 on visibility impairment in Beijing, China. Aerosol & Air Quality Research, 2014, 14(1): 260-268.
[7] 杜荣光, 齐冰, 胡德云, 等. 杭州市区相对湿度及PM2.5对能见度的影响分析. 南京大学学报(自然科学), 2015(3): 473-480.
[8] 王英, 李令军, 李成才. 北京大气能见度和消光特性变化规律及影响因素. 中国环境科学, 2015, 35(5): 1310-1318.
[9] Chen J, Xin J, An J, et al. Observation of aerosol optical properties and particulate pollution at background station in the Pearl River Delta region. Atmospheric Research, 2014, 143: 216-227.
[10] 白永清, 祁海霞, 刘琳, 等. 武汉大气能见度与PM_(2.5)浓度及相对湿度关系的非线性分析及能见度预报. 气象学报, 2016, 74(2): 189-199.
[11] Brock CA, Wagner NL, Anderson BE, et al. Aerosol optical properties in the southeastern United States in summer - Part 1: hygroscopic growth. Atmospheric Chemistry & Physics, 2016, 15(18): 25695-25738.
[12] 樊高峰, 马浩, 张小伟, 等. 相对湿度和 PM2.5浓度对大气能见度的影响研究: 基于小时资料的多站对比分析. 气象学报, 2016, 74(6): 959-973.
[13] 王淑英, 张小玲, 徐晓峰. 北京地区大气能见度变化规律及影响因子统计分析. 气象科技, 2003, 31(2): 109-114.
[14] Lv B, Hu Y, Chang H H, et al. Daily estimation of ground-level PM2.5 concentrations at 4 km resolution over Beijing-Tianjin-Hebei by fusing MODIS AOD and ground observations. Science of the Total Environment, 2016, 580: 235.
[15] Lowenthal D H, Kumar N. PM2.5 mass and light extinction reconstruction in IMPROVE. Journal of the Air & Waste Management Association, 2003, 53(9): 1109.
[16] Pitchford M, Maim W, Schichtel B, et al. Revised algorithm for estimating light extinction from IMPROVE particle speciation data. Journal of the Air & Waste Management Association, 2007, 57(11): 1326.
[17] Wang J L, Zhang Y H, Shao X L, et al. Quantitative relationship between visibility and mass concentration of PM2.5 in Beijing. 环境科学学报(英文版), 2006, 18(3): 475-481.
[18] 吴兑, 刘啟汉, 梁延刚, 等. 粤港细粒子(PM2.5)污染导致能见度下降与灰霾天气形成的研究. 环境科学学报, 2012, 32(11): 2660-2669.
[19] Zhu L, Zhu G, Han L, et al. The Application of Deep Learning in Airport Visibility Forecast. Atmospheric & Climate Sciences, 2017, 7(3): 314-322.
[20] Byun D, Schere K L. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Applied Mechanics Reviews, 2006, 59(2): 51-77.
[21] 傅刚, 李晓岚, 魏娜. 大气能见度研究. 中国海洋大学学报(自然科学版)自然科学版, 2009, 39(5): 855-862.
[22] Lyu B, Hu Y, Sun X, et al. A fusion method combining ground-level observations with chemical transport model predictions using ensemble deep learning framework: application in China estimating spatiotemporally-resolved PM2.5 exposure trends 2014-2017. Environmental Science & Technology, 2018, to be submitted.
[23] Barth A, Azcárate A A, Joassin P, et al. Introduction to optimal interpolation and variational analysis. SESAME Summer School, 2008.
[24] Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks, 2015, 61: 85-117.
[25] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information & Computer Sciences, 2003, 43(6): 1947.
[26] Ridgeway G. Generalized boosted models: a guide to the GBM package. https://cran.r-project.org/web/packages/gbm/gbm.pdf.
[27] van der Laan M J, Polley E C, Hubbard A E. Super learner. Statistical Applications in Genetics & Molecular Biology, 2007, 6(1): Article25.
[28] Koch S E, Desjardins M, Kocin P J. An interactive Barnes objective map analysis scheme for use with satellite and conventional data. Journal of Applied Meteorology, 1983, 22(9): 1487-1503.
[29] 王叶红, 赵玉春, 崔春光. 多普勒雷达估算降水和反演风在不同初值方案下对降水预报影响的数值研究. 气象学报, 2006, 64(4): 485-499.
[30] Ma Z, Hu X, Sayer A M, et al. Satellite-based spatiotemporal trends in PM2.5 concentrations: China, 2004–2013. Environmental Health Perspectives, 2015, 124(2): 184-192.

X波段全固态双线偏振一体化天气雷达

刘强 苗雷

X波段固态天气雷达

X波段全固态双线偏振一体化天气雷达专为气象部门用户设计，主要应用于人工影响天气和重点区域组网监测等领域。该雷达采用全固态发射机、双通道恒温接收机和一体化天伺系统等先进技术系统，具有高效率、高机动性和高稳定性的特点。该雷达能够定量估测降水、识别降水粒子相态，有助于了解降水微物理结构，带来更多、更全面的分析资料，并可对灾害天气进行自动报警，提高气象保障能力。

（作者单位：北京敏视达雷达有限公司）