

安徽省数字气象档案馆关键技术研究与应用

■ 江双五 孔芹芹 谢伟 刘惠兰 华连生

围绕安徽省数字气象档案馆业务能力提升,重点从“人一馆一物”一体化的智能物联技术、气象档案知识图谱和智能检索技术、数字气象档案资源自动归档及长期存储技术三个方面,系统总结和介绍了安徽省数字气象档案馆创新性应用成果,提出了深化数字气象档案馆建设的可能发展方向。安徽省数字气象档案馆关键技术集成应用是信息技术提升气象档案业务管理能力的一次有益探索,对开展全国数字气象档案馆建设具有一定的参考价值。

中图分类号: P4

文献标志码: A

DOI: 10.3969/j.issn.2095-1973.2024.05.010

安徽省气象档案馆是专业性档案馆,馆藏档案以纸质载体和数字载体为主,保存的纸质气象记录档案超过千万页,来源于电子文件和纸质档案数字化文件的数字档案达30 TB,种类涵盖地面、高空和雷达等气象记录档案的大部分类别,另外每年有超过3 TB的电子档案需要归档。实现气象档案资源快速收集和整理、安全高效保存,为用户提供好用的档案资源和搭建好档案利用的平台,是安徽省数字气象档案馆近几年建设的主要内容。

在21世纪初档案部门就提出了数字档案馆建设的设想和发展目标。经过二十多年的发展,数字档案馆从概念提出到初步实践,从理论探讨到技术应用,从各省多头探索到国家统一规范,经历了初步探索、普遍应用到深化发展三个阶段。2018年以来,中国气象局为有效提升气象档案服务于行业内外能力,尤其是服务一般用户(气象爱好者、在校学生等)的能力,顺应数字档案馆发展趋势提出了建设数字气象档案馆的目标,成立数字气象档案馆建设技术组,统筹安排安徽、山东、河北和湖北等省气象局共同开展数字气象档案馆建设试点工作。技术组面向档案安全应用的需要,根据数字气象档案馆建设的难点和重点,聚焦档案收集、保管和利用,提出了“人一馆一物”一体化的智能物联技术、数字气象档案资源自动归档及长期存储技术、气象档案知识图谱构建和智能检索技术,作为数字气象档案馆建设的三大关键技术。经过几年的建设,三大关键技术取得重要进展并在全国气象档案馆建设中逐步推广应用,气象档案的安全高效

管理和服务能力得到了重要提升。

1 “人一馆一物”一体化的智能物联技术

“人一馆一物”一体化管理的“人”为档案管理员和借阅人员,“馆”为档案馆环境,“物”为档案。利用物联网技术构建的该种管理模式可以提升档案进出馆安全管理、保管环境的标准化调控、档案盘点和借阅的自动化和智能化能力,有效改变传统手工式的气象档案服务和管理模式。随着物联网技术不断深入发展,该技术深刻地影响或改变着档案管理和服务模式。安徽省数字气象档案馆智能物联系统应用技术架构主要由感知层、控制层、应用层和传输层组成(图1)。

感知层主要负责档案馆中各种环境、安防和实体管理(RFID)信息的采集。应用层主要实现环境调控和视频监控设备的运行、门禁的开关、信息检索和展示、自动盘点和语音播报等功能。控制层通过传输层接收感知层信息,依据预定的规则分析判断,向应用层设备发布状态调整指令。各层之间分工合作,控制环境控制设备的运行,使馆藏环境符合《数字气象档案馆建设规范》《气象档案馆环境监测技术规范》等中国气象局气象信息化系统工程相关标准规定要求。

1.1 气象档案馆安防管理技术

安防管理主要实现档案馆的安全运行,防止未经授权的人员闯入。该技术主要通过人脸识别设备、门禁控制、视频监控、告警和展示等设备的协同工作来完成。人员进入档案馆时需要通过人脸识别设备

收稿日期: 2024年1月31日; 修回日期: 2024年4月18日

第一作者: 江双五(1971—), Email: jiangjiang0621@sina.com

资助信息: 国家档案局科技项目(2022-X-060); 中国气象局档案建设专项(YBSZX2024007)

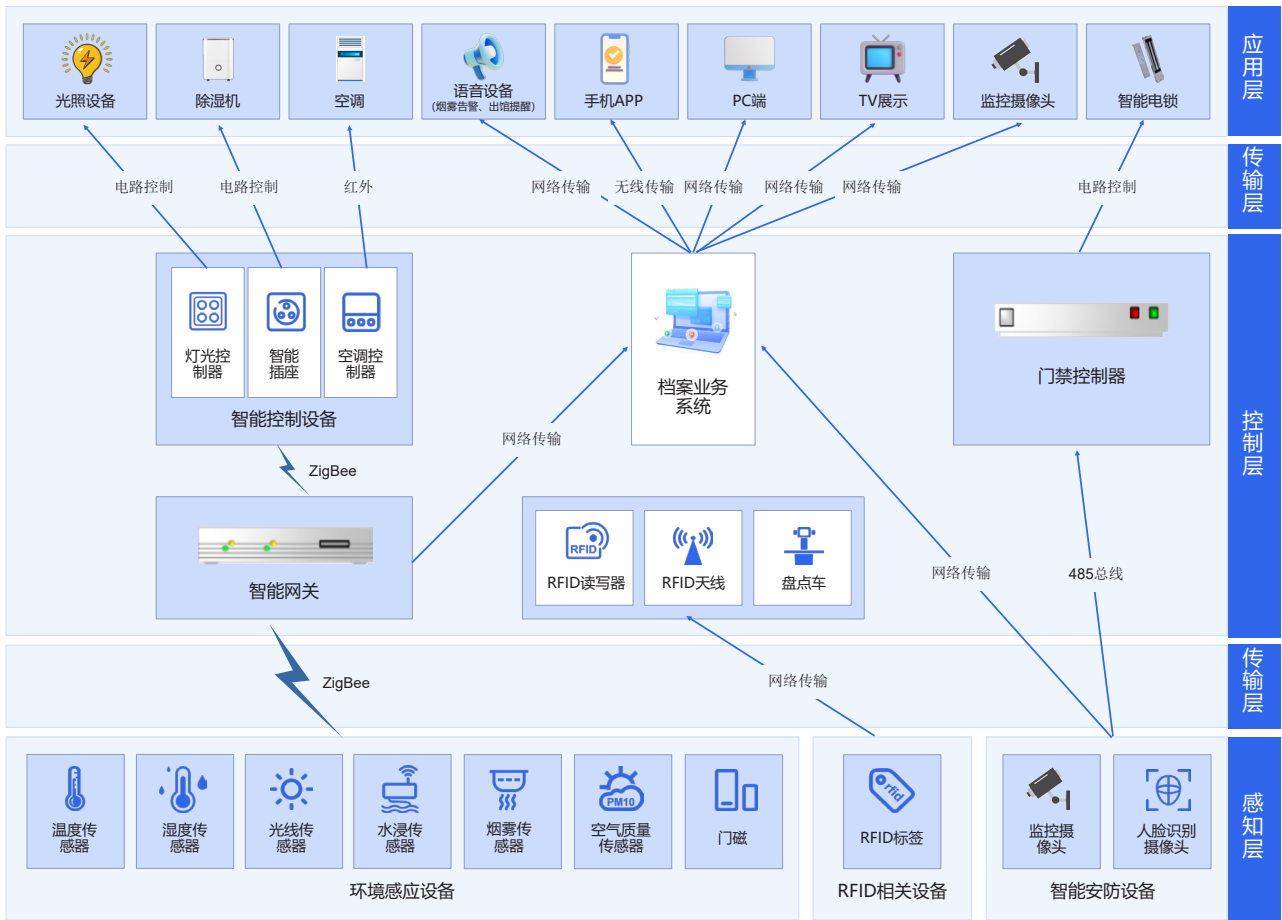


图1 智能物联网系统架构

验证，已授权可直接进入，非授权用户门禁将不能打开，系统提示用户需要注册登记，视频监控设备将对进馆情况自动拍照留存。当用户强行进入档案馆时，入侵报警模块可提供短信、语音和视频抓拍来实现报警。

1.2 气象档案馆环境管理技术

环境信息的管理实现档案馆环境信息的自动感知和控制。该技术模块由环境感知、指挥和信息展示三部分组成。通过环境感知传感器感知档案馆的温度、湿度、空气质量、灯光和水浸和烟感等，将感知的数据通过ZigBee技术汇集到物联网网关协调器，同时网关协调器通过以太网与档案业务系统中指挥模块交互。指挥模块分析环境信息，当环境信息异常时可依据环境问题发出相应指令触发空调、除湿机、加湿机、灯光控制等环境调节设备的运行，同时将告警信息方式通知档案管理员。信息展示等终端可调用档案业务系统接口实现对馆区环境的温度、湿度、光强、空气质量等进行监测，并可接收漏水、烟雾等告警信息。

1.3 气象档案馆实体档案智能管理技术

实体档案智能管理主要提升实体档案的借阅、归还、盘点等管理水平。该技术模块主要由RFID模块和档案管理模块组成。RFID模块包括RFID标签、读写器、天线、工作台和盘点车等。其中RFID标签粘贴在档案上，RFID读写器和天线用馈线相连装在档案馆内的天花板或者地面上。档案管理模块主要包括档案借阅归还出入管理、盘点和标签质量检查等模块。

2 气象档案知识图谱构建和智能检索技术

数字气象档案馆一项基本任务是为用户提供好用的档案并搭建好档案利用的平台。气象档案知识图谱构建和智能检索技术就是解决档案利用的一项关键技术，根据应用场景快速聚集关联档案资源，通过智能检索平台让不熟悉气象档案一般用户可方便快捷地从专业档案中找到自己需要的信息。气象档案知识图谱构建及智能检索技术层次架构主要由气象档案的收集整理、知识建模、词库的构建、知识抽取、知识融合、图谱数据存储、知识图谱应用7层构成(图2)。

收集整理的资料有三种：非结构化纸质数字化资

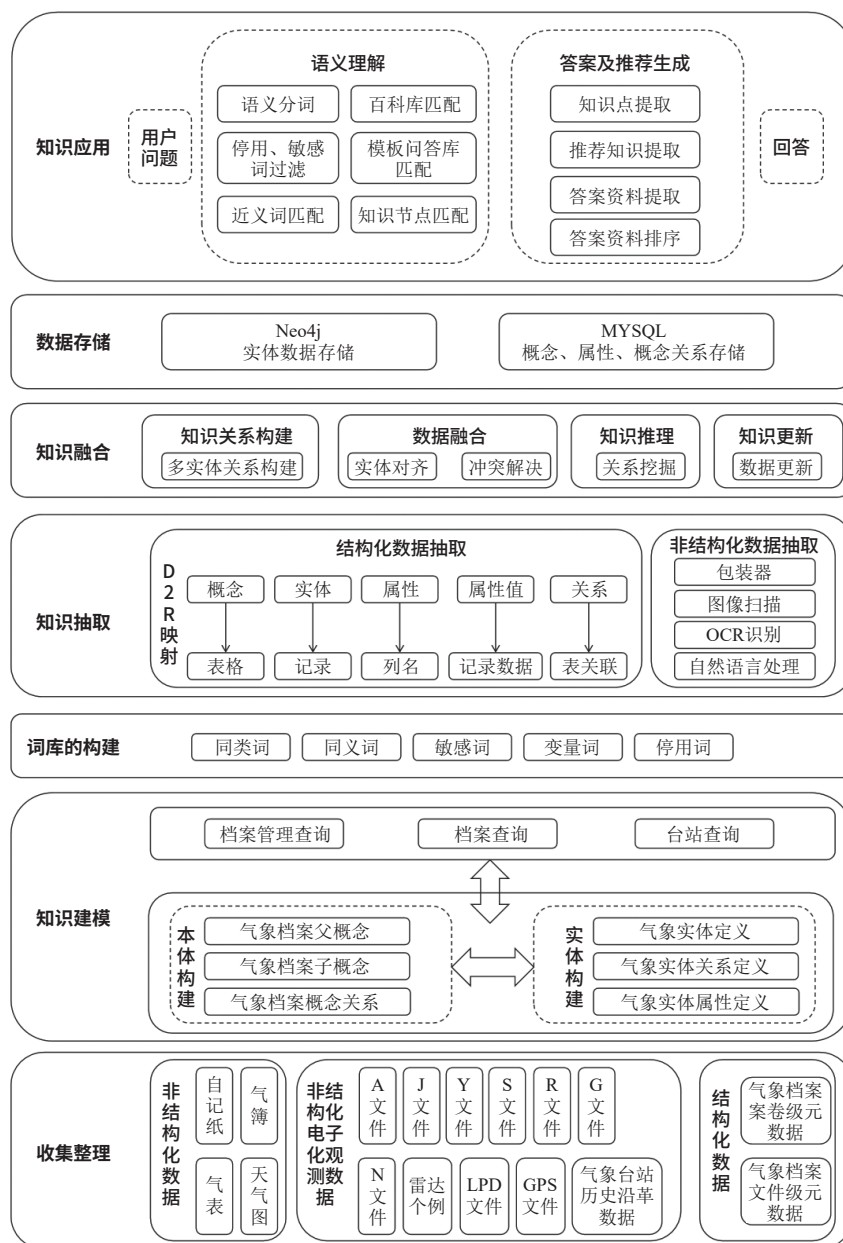


图2 气象档案智能检索系统架构

料、非结构化原生电子观测数据、结构化数据。非结构化纸质数字化资料包含自记纸、气表、气簿、天气图等档案。非结构化原生电子观测数据包含A文件、J文件、Y文件、S文件、R文件、G文件、N文件、雷达个例、LPD文件、GPS文件、气象台站历史沿革数据。结构化数据主要为气象档案元数据，包含气象档案案卷级元数据和文件级元数据。

2.1 基于应用场景气象记录档案知识图谱模型的构建

气象档案知识图谱模型需要根据应用场景的需求来设计。总结用户进馆查询档案的应用场景，分为三大应用场景：查档案、查台站和查要素。以知识图

谱的形式构建专题资源模型，快速聚集应用场景查询需求，模型具体构建流程为梳理分析整合、加工后的“台站、档案、要素”资源，枚举“台站、档案、要素”相关所有基本术语及相关主题词，用本体构建技术，经过气象领域专家的指导，定义“台站、档案、要素”专题领域本体概念，采用自顶向下的方式定义概念层次结构，然后定义不同层次的概念属性、属性约束、数值范围等，最后定义概念与概念、属性与属性之间的关联关系，从而构建“台站、档案、要素”本体模型，并形成“台站、档案、要素”本体知识库。随着专题档案资源的不断扩充，也会有新的概念、属性、关系诞生，这就需要对本体模型进行补充

和完善。在知识图谱中，实体是最基本的元素，实体代表客观存在的事物，如黄山站就是一个实体，它是知识图谱中“图”的一个节点，实体间的关系是“图”中节点之间的连线。本体概念是对实体的一个抽象描述，它们之间存在对应关系，为本体概念添加实例通

常被称为本体填充。本体填充是在本体模型构建完成之后进行的，首先进行概念、概念属性及概念关系的定义，然后为定义好的概念及概念关系添加实例，形成气象记录档案实体关系模型。图3为黄山站变迁本体实体关系映射关系图。

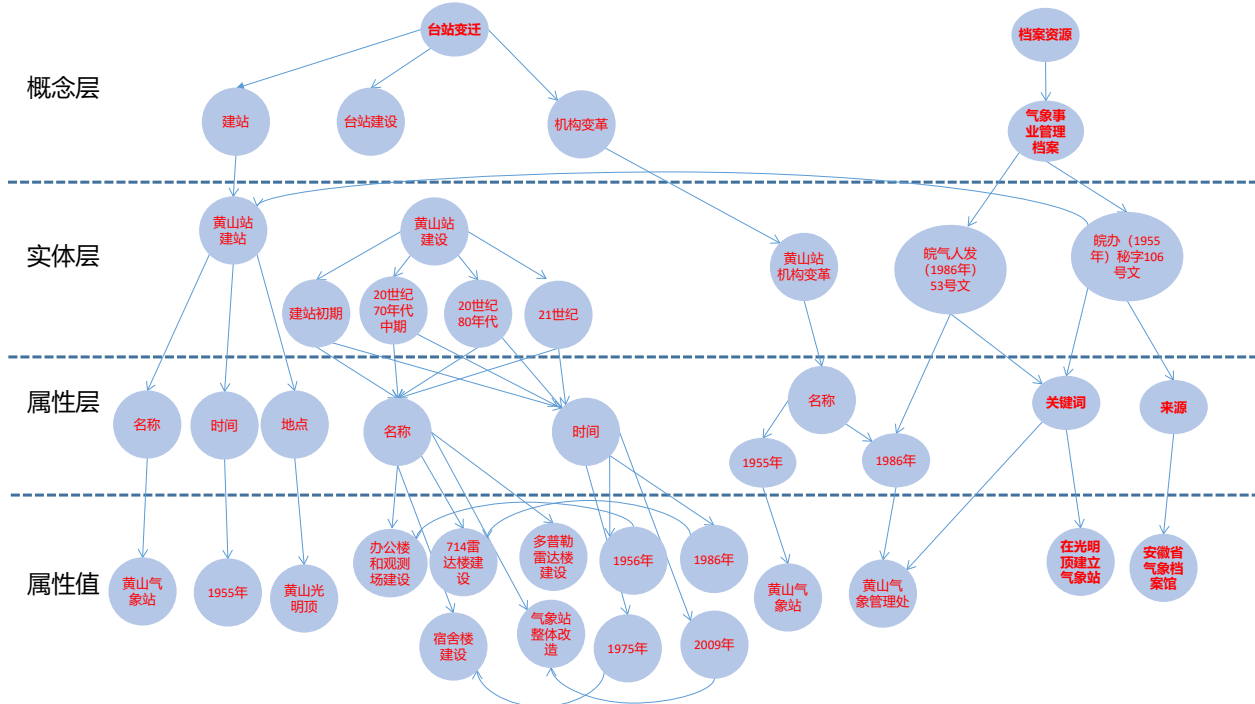


图3 黄山站变迁本体实体关系映射关系图

2.2 词库的构建

词库由大量词集构成。词集分为同类词、同义词、敏感词、变量词、停用词、短语法和知识问答库等。词集种类和内容越丰富，匹配算法等效果越好，能极大地提高问答回复的精准度。

不同词库采用不同构建方法。例如，同义词库的构建，一方面，基于沿革数据中的变革项，抽取各时间点的沿革数据，构建同义词库；另一方面，依托索引分词库，采用文本相似度算法，设定相似度阈值，符合条件的词可加入同义词库中；同时，为领域用户提供同义词维护功能，依托业务积累，延伸出同义词，添加到同义词库中。构建同义词库能帮助系统理解用户的检索意图，提高全文检索的准确率，为用户返回想要的结果。

解决气象档案快速检索问题，最有效的方案是建立知识问答库。气象档案的知识问答基本上属于人、事、物、台站、气象要素、时间概念（实体）范畴，高频问题主要来源于这些概念或实体的组合。气象记录档案的知识问题具有一定的可预知性，预设高频问

题可提高知识问答的准确度和查询速度。目前安徽省气象档案业务系统已建立了智能问答模板超过32242个，台站名称方面82个，档案概念23个，档案内容方面48个，档案要素方面112个，台站属性方面16个，档案属性方面13个，台站和台站属性组合1312个、档案和档案属性方面624个，台站与档案概念组合问题1886个，台站与档案组合问题3936个，台站与观测要素组合问题9184个，台站档案概念时间组合问题超过1886个，台站档案时间组合问题超过3936个，台站要素时间组合超过9184个。

2.3 智能检索应用场景展示

建立气象档案知识图谱的目的是为用户提供好用的气象档案知识。安徽省数字气象档案馆按照应用场景设计了两大类展示系统，分别为台站相关的智能检索展示和档案相关的智能检索展示。

2.3.1 台站相关智能检索展示设计

基于台站相关知识图谱模型以及知识问答库等设计台站相关的智能检索展示系统。例如，档案利用人员登录，检索框输入需要查询的台站名称，如“黄

山站”，点击查询按钮后，跳转至查询结果显示页面。页面左侧上方显示被查询台站百科信息；台站百科下方显示台站的“大事记模块”，用户查询该站历史上特定重大灾害性事件时，可看到该灾害发生前后各类气象观测记录档案，受灾情况和救灾情况相关新闻报道、预警信息发布等文书档案；台站“大事记模块”下方显示台站的“气候变迁”，点击后可查看该站建站以来温度、降水等主要观测要素年尺度变化情况；“气候变迁”下方显示检索词分词检索后的档案列表。页面右侧上方显示“气象观测类档案统计”，为被检索台站的相关气象观测类档案；页面右侧中间显示“其他类档案推荐”，为被检索台站相关的非气象观测类档案；页面右侧下方显示“台站推荐”，为与被检索台站同市的其他台站；页面右侧最下方显示“知识图谱”。

2.3.2 档案相关智能检索展示设计

基于档案相关知识图谱模型以及知识问答库等设计档案相关的智能检索展示系统。例如，档案利用人员登录，检索框输入需要查询的档案子分类名称，如“气簿1”，点击查询后，跳转至查询结果显示页面，页面左侧上方显示被查询档案分类的百科，观测时间沿革信息下方显示检索词分词检索后的档案列表；页面右侧显示“台站资源统计”，所有含有被检索类型档案的台站推荐；页面右侧下方显示“相关推荐、知识图谱”。

3 数字气象档案资源自动归档及长期存储技术

为用户提供好用的档案资源，首先数字气象档案馆要有充足的能长期（30年以上）或永久安全保管档案资源的能力。相对于传统档案以实物档案为主，数字档案馆更注重电子档案的保管。解决海量气象数据自动归档和长期存储难题，是数字气象档案馆另一项关键技术问题。根据实际业务需求，数字气象档案自动归档及长期保存子系统业务架构包括数据来源层、数据存储层、作业调度层、业务管理层、应用层。

数据来源层主要完成待归档资料的收集。安徽省自动归档资料主要包括省级“天擎”系统和各单位移交的气象电子档案。待归档的气象电子档案上传到文件服务器（NAS存储），通过采集归档调度任务，周期性地从文件服务器上采集待归档的气象电子档案。

数据存储层主要实现归档资料的自动整理、自动提取、自动打包、自动刻录和长期保存任务。存储设施主要包括数据库服务器、文件服务器和长期存储设备。数据库服务器主要保存归档元数据、归档管理运

行数据。文件服务器主要保存临时存储待归档气象观测数据、归档整理文件、打包存储文件、周期性清理文件等。存储设备包括存储载体以及存储载体的装具设备和读取设备，存储载体包括固态硬盘（缓存）、磁盘（在线存储备份）、蓝光光盘（离线长期保存）。在线备份/近线备份的设备为光磁一体机、离线备份设备为防磁柜或者离线光盘库等。

作业调度层主要完成自动归档长期存储系统运行中的作业调度，主要包括采集归档、存储刻录、结果查询和定时清理等作业调度。通过作业调度任务的执行，完成自动归档及长期保存的全流程批量数据处理等。

业务管理层主要完成自动归档和长期存储业务运行中的各种作业管理，主要包括归档元数据、策略（采集归档策略、存储刻录策略、定期清理策略）、定期检测、归档统计、任务调度、预警、日志和目录管理等，由省级档案管理员负责省级归档全流程的业务管控。

业务应用层为档案管理员或用户提供档案利用功能，主要功能为在气象档案业务系统（云历）提供归档数据检索、电子档案借阅（在线浏览、下载、复制）等在线利用功能。

4 总结与展望

从2018年承担中国气象局数字气象档案馆试点任务以来，安徽省数字气象档案馆建设已开展了六年。随着智能物联、基于知识图谱的智能检索和数字档案资源长期存储等关键技术安徽省气象档案馆的逐步研发和应用，安徽省数字气象档案馆自动化和智能化水平逐步提升，已基本建成场馆环境管理规范化、实体管理物联化、档案收集网络化、档案存储安全化、档案服务便捷化和场景化，为档案利用人员提供智能和高效的档案服务平台，为档案管理员和职能管理部门提供规范、安全和高效的档案管理平台。

随着海量气象资料的逐年增加以及大模型为代表的人工智能技术的深入发展，安徽省数字气象档案馆仍需要从以下三个方面进一步完善：一是深化数字气象档案资源长期存储系统建设，优化电子档案长期存储管理策略，实现多套长期存储设备自动互备，自动检测和应急恢复功能；二是开展大模型为代表的人工智能技术在气象档案系统建设中的深入应用，实现气象档案利用的高效和智能；三是气象档案编研产品的研发及应用。通过特定事件档案服务应用场景的搭建、海量信息提取与分析技术，实现档案结构化与非结构化数据编研分析及产品定制，为气象档案用户提

供精准的、高质量的档案溯源信息服务，真正体现档案的价值和作用。

致谢：感谢全国数字气象档案技术组对安徽省数字气象档案馆关键技术研究的大力支持。

深入阅读

蔡学美, 2014. 数字档案馆发展阶段和建设内容[J]. 档案管理(1): 26-27.
曹明宇, 李青青, 杨志豪, 等, 2019. 基于知识图谱的原发性肝癌知识问答系统[J]. 中文信息学报, 33(6): 88-93.
江双五, 温华洋, 盛绍学, 等, 2021. 智慧档案馆背景下安徽省气象档案业务系统[J]. 计算机系统应用, 30(9): 128-137.
江双五, 刘惠兰, 温华洋, 等, 2022. 气象记录档案知识图谱构建[J]. 计算机系统应用, 31(1): 73-82.
江双五, 刘惠兰, 华连生, 等, 2023. 基于知识图谱技术的安徽省数字气象档案馆智能物联系统设计[J]. 长江信息通信, 36(9):

132-136.
蒋建峰, 金怿, 2019. 智慧档案馆建设的实践与探索[J]. 档案与建设(2): 46-50.
李贺, 刘嘉宇, 李世钰, 等, 2021. 基于疾病知识图谱的自动问答系统优化研究[J]. 数据分析与知识发现, 5(5): 115-126.
田雷, 2015. 物联网技术在智慧档案馆建设中的应用研究[J]. 档案学通讯(1): 60-64.
杨靖, 朋礼青, 2019. 人工智能对智慧档案馆的驱动作用研究[J]. 北京档案(1): 9-13.

(作者单位: 安徽省气象信息中心)

(编辑: 卢冰)

(上接28页)

- [51] Xie J B, Zhang M H, Xie Z H, et al. An orographic-drag parametrization scheme including orographic anisotropy for all flow directions[J]. Journal of Advances in Modeling Earth Systems, 2020, 12(3): e2019MS001921.
- [52] Dan J Y, Gao Y H, Zhang M. Detecting and attributing evapotranspiration deviations using dynamical downscaling and convection-permitting modeling over the Tibetan Plateau[J]. Water, 2021, 13(15): 2096.
- [53] Liu Z Y, Gao Y H, Zhang G. How well can a convection-permitting-modelling improve the simulation of summer precipitation diurnal cycle over the Tibetan Plateau?[J]. Climate Dynamics, 2022, 58(11-12): 3121-3138.
- [54] Gao Y H, Chen F, Jiang Y S. Evaluation of a convection-permitting modeling of precipitation over the Tibetan Plateau and its influences on the simulation of snow-cover fraction[J]. Journal of Hydrometeorology, 2020, 21(7): 1531-1548.
- [55] Basist A, Bell G D, Meentemeyer V. Statistical relationships between topography and precipitation patterns[J]. Journal of Climate, 1994, 7(9): 1305-1315.
- [56] 曾礼, 高艳红, 蒋盈沙, 等. 地形因子对华东地区降水影响的尺度效应研究[J]. 地球科学进展, 2022, 37(5): 535-548.
- [57] Wang Y. The upper airflow prediction by PLS (Partial Least Square) model[C]//The Scientific Conference on the South China Sea Monsoon Experiment (SCSMEX). Shanghai, 2001.
- [58] 舒守娟, 王元, 熊安元. 中国区域地理、地形因子对降水分布影响的估算和分析[J]. 地球物理学报, 2007, 50(6): 1703-1712.
- [59] Brunson C, Fotheringham A S, Charlton M E. Geographically weighted regression: a method for exploring spatial nonstationarity[J]. Geographical Analysis, 1996, 28(4): 281-298.
- [60] 杨顺华. 基于空间回归模型的土壤有机质区域分布特征研究[D]. 武汉: 华中农业大学, 2016.
- [61] Kumar S, Lal R, Liu D S. A geographically weighted regression kriging approach for mapping soil organic carbon stock[J]. Geoderma, 2012, 189-190: 627-634.
- [62] Fotheringham A S, Yang W B, Kang W. Multiscale geographically weighted regression (MGWR)[J]. Annals of the American Association of Geographers, 2017, 107(6): 1247-1265.
- [63] 王璐, 沈学顺. 对流尺度集合预报与模式不确定性研究进展[J]. 气象, 2019, 45(8): 1158-1168.
- [64] Eliassen A. The quasi-static equations of motion with pressure as independent variable Geofys[J]. Publikasjoner, 1949, 17(3): 1-44.
- [65] 宇如聪, 曾庆存, 彭贵康, 等. “雅安天漏”研究II. 数值预报试验[J]. 大气科学, 1994, 18(5): 535-551.
- [66] 程锐, 宇如聪, 徐幼平, 等. 一个基于Eta垂直坐标的新WRF动力框架及其数值试验[J]. 大气科学, 2022, 46(2): 237-250.
- [67] 胡江林, 王盘兴. 地形跟随坐标下的中尺度模式气压梯度力计算误差分析及其改进方案[J]. 大气科学, 2007, 31(1): 109-118.
- [68] 张朝林, 季崇萍, Kuo Y H, 等. 地形对“00.7”北京特大暴雨过程影响的数值研究[J]. 自然科学进展, 2005, 15(5): 572-578.
- [69] Duda J D, Wang X G, Xue M. Sensitivity of convection-allowing forecasts to land surface model perturbations and implications for ensemble design[J]. Monthly Weather Review, 2017, 145(5): 2001-2025.
- [70] Martin T C M, Rocha H R, Perez G M P. Fine scale surface climate in complex terrain using machine learning[J]. International Journal of Climatology, 2021, 41(1): 233-250.

(编辑: 郑秋红)