

# 数据驱动的定量遥感研究进展与挑战

杨倩倩<sup>1</sup>, 靳才溢<sup>1</sup>, 李同文<sup>2</sup>, 袁强强<sup>1</sup>, 沈焕锋<sup>3</sup>, 张良培<sup>4</sup>

1. 武汉大学 测绘学院, 武汉 430079;

2. 中山大学 测绘科学与技术学院, 珠海 519082;

3. 武汉大学 资源与环境科学学院, 武汉 430079;

4. 武汉大学 测绘遥感信息工程国家重点实验室, 武汉 430079

**摘要:** 定量遥感是从原始遥感观测信息中定量推算或反演出地学参量的理论与方法。传统定量遥感主要基于模型驱动, 强调通过数学或物理模型完成推算和反演。随着人工智能技术的发展和普及, 数据驱动的方式也逐渐受到广泛关注, 其强调的是通过机器学习等方式挖掘遥感观测数据中所包含的信息, 完成地学参量的定量反演。在强大计算能力的支持下, 数据驱动的方法在定量遥感的多个领域都取得了可喜的成就, 但其也具有弱化物理规律及因果关系的缺点。在此背景下, 耦合物理规律和机器学习, 发展模型和数据共同驱动的反演框架, 开始成为新的研究热点。以机器学习辅助物理模型, 或以物理规律约束机器学习的研究逐步展开, 并初见成效。然而, 仍面临着较大的挑战, 联合模型的不确定性、泛化性、可迁移性, 以及小样本情况下的联合建模等都是急需解决的问题。未来, 数据驱动和模型驱动的深度耦合将是待突破的难点和重点。

**关键词:** 遥感, 定量遥感, 模型驱动, 数据驱动, 深度学习, 融合

**引用格式:** 杨倩倩, 靳才溢, 李同文, 袁强强, 沈焕锋, 张良培. 2022. 数据驱动的定量遥感研究进展与挑战. 遥感学报, 26(2): 268-285

Yang Q Q, Jin C Y, Li T W, Yuan Q Q, Shen H F and Zhang L P. 2022. Research progress and challenges of data-driven quantitative remote sensing. National Remote Sensing Bulletin, 26(2): 268-285 [DOI:10.11834/jrs.20211410]

## 1 引言

遥感卫星可以对地球表面进行大范围、持续的观测, 为国家宏观决策、资源调查、灾害监测等提供及时、有效的数据支撑(徐冠华, 1997)。遥感卫星获取的原始数据主要是电磁波信号, 需要经过一系列的转换才能变成易理解、可应用的地表参量信息, 这个转换过程被称为定量遥感。具体来说, 定量遥感是利用遥感传感器获取的地表地物的电磁波信息, 在先验知识和计算机系统支持下, 定量获取观测目标参量或特性的方法与技术(李小文, 2005)。随着遥感科学的不断发展, 越来越多的卫星被发射升空, 可获取的卫星数据类型变得复杂多样, 数量也急剧增加, 为定量遥感提出了新的挑战(宫鹏, 2009; 梁顺林等, 2013, 2016)。

作为新兴的遥感信息获取与分析方法, 定量遥感强调通过数学的或物理的模型将遥感信息与观测地表目标参量联系起来, 定量地反演或推算出某些地学目标参量(李小文, 2005)。经典基于物理的定量遥感参数反演建模, 依赖于对遥感系统机理、辐射传输理论、地表参量光学、几何等特征的理解(图1)。例如, 在植被领域, Jacquemoud和Baret(1990)根据叶片内部辐射传输机制, 建立了叶片的反射率/透过率估算模型PROSPECT; 而后, Gastellu-Etchegorry等(1996)基于三维辐射传输理论, 建立了植被冠层模拟模型DART, 实现了利用数学模型从遥感信号中提取植被信息的目的; 近年来, Van der Tol等(2009)结合辐射传输和能量守恒, 发展出了用于日光诱导叶绿素荧光模拟的SCOPE模型。在水文领域, 以地表能量平衡为基础的蒸散发估算模型Penman-Monteith

收稿日期: 2021-06-15; 预印本: 2021-10-20

基金项目: 国家自然科学基金(编号:41922008); 中国科学院战略性先导专项(编号:XDA19090104)

第一作者简介: 杨倩倩, 1995年生, 女, 博士研究生, 研究方向为大气遥感。E-mail: qianqian\_yang@whu.edu.cn

通信作者简介: 袁强强, 1985年生, 男, 教授, 研究方向为遥感信息处理, 定量遥感, 机器学习。E-mail: qqyuan@sgg.whu.edu.cn

(Monteith, 1965) 和 Priestley-Taylor 公式 (Priestley 和 Taylor, 1972), 以及考虑积雪几何结构及微波辐射传输理论的雪参数计算模型 Bicontinuous (Zhu 等, 2018), 都得到了广泛的应用。大气领域, 根据光谱吸收特征发展而来的差分吸收技术 (陈良富 等, 2015) 是目前痕量气体反演的主流方法; Koelemeijer 等 (2006) 基于气溶胶基本物理属性构建的从遥感气溶胶光学厚度 AOD (Aerosol Optical Depth) 估算大气颗粒物质量浓度的理论公式, 奠定了模型驱动的大气颗粒物浓度遥感反演的物理基础。这些方法有一个共同的特点, 即都是从物理规律出发, 构建从观测数据到目标参量之间的表达式, 称之为模型驱动的方法。这类方法的优点是, 观测数据和目标参量之间具有明确的因果关系, 模型建立过程透明, 可解释。但是, 物理建模的方式难以准确地描述复杂多变的现实情况, 一个精确、逼真的模型往往又十分复杂, 计算成本高昂。除此之外, 物理模型包含大量变量, 这些变量难以获取和计算, 也会给建模带来巨大的不确定性。

近年来, 随着人工智能技术的发展和大数据时代的到来, 数据驱动的定量反演方法开始飞速

发展。如图 2 所示, 这类方法从初始的数据或观测值出发, 利用机器学习等工具寻找和建立内部特征之间的关系, 从而实现时空预测或发现物理规律。得益于计算能力和人工智能算法的飞速发展, 基于数据驱动的定量遥感反演方法在效率、精度等多方面都取得了卓越的成就, 越来越广泛地被应用于植被、水文、冰冻圈、大气、地质等多个领域 (陈仲新 等, 2016; 李德仁 等, 2014b; Yuan 等, 2020; 张兵, 2018; Reichstein 等, 2019; Tuia 等, 2021)。然而, 数据驱动的方法重在数据挖掘, 忽视物理规律, 因此, 往往无法解释因果关系。这种机理的不透明性、因果的不可解释性也使得其受到了一定的争议和质疑。

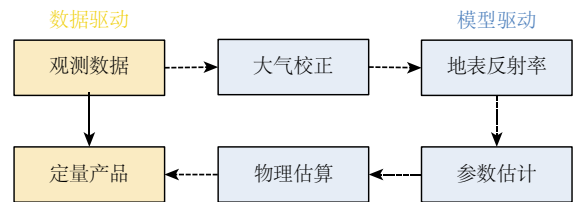


图 1 模型驱动和数据驱动的定量遥感原理示意图  
Fig. 1 Schematic diagram of the principles of model-driven and data-driven quantitative remote sensing

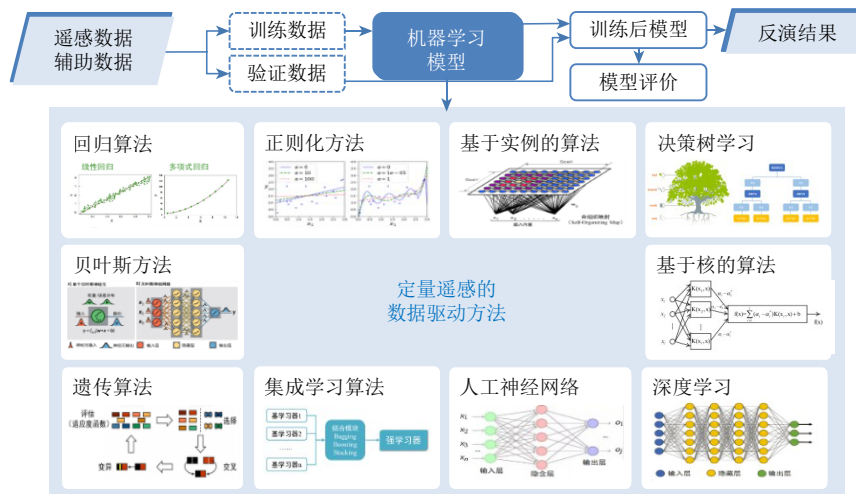


图 2 定量遥感数据驱动方法流程示意图及方法分类概述  
Fig. 2 Flowchart and classification of data-driven

模型驱动和数据驱动的方法优势互补, 思考如何将二者结合, 进一步提升定量遥感反演的精度, 将是我们所面临的新一轮挑战。本文试图针对数据驱动的定量遥感方法进行概括, 对现有的数据和模型驱动相结合的方法进行了总结, 并梳理了目前我们所面临的挑战和一些潜在的发展方向。

## 2 数据驱动的定量遥感主流方法

本文列举了数据驱动的定量遥感反演中常用的机器学习算法, 并根据算法的功能和形式的相似性, 把算法进行了简单的分类 (图 2)。针对每一类机器学习算法, 首先简单介绍其原理, 阐述其在定量遥感中典型的应用实例, 最后总结了该类方法的用于定量反演的优缺点 (表 1)。

表1 数据驱动的定量遥感方法原理及特点总结

Table 1 A summary of the principles and characteristics of data-driven quantitative remote sensing retrieval methods

类别	原理	代表算法	特点
回归算法	假定目标值与特征之间的关系,通过构建损失函数,来求解损失函数最小时的参数	线性回归	简单,高效,稳定; 是许多强大的非线性模型的基础; 易于理解,结果具有较好的可解释性; 对于数据量小、关系简单的数据很有效; 难以很好地表达复杂数据及非线性关系
		多项式回归	比线性回归更灵活; 可以模拟一些复杂的关系
		逐步式回归	可以选择和控制输入变量,取得最优解释变量集; 得到的最优变量集有一定的随机性和不可解释性
		多元自适应回归样条	能够处理数据量大、维度高的数据; 且计算快捷、模型精确,可解释性强
正则化方法	向模型中引入附加约束以防止过拟合和提升模型性能的一类方法	岭回归, Lasso 回归, 弹性网络	在输入变量具有高维度和多重共线性的情况下也能有良好的效果; 可适用于小样本数据
基于实例的算法	选取一批样本数据,然后根据某些近似性把新数据与样本数据进行比较,实现回归或分类	K 邻近回归, 自组织映射算法	内存使用往往较高、计算成本高; 无法用于高维特征空间
决策树学习	根据数据的属性采用树状结构建立的决策模型	ID3, CART, C4.5	单棵决策树的学习能力有限; 容易过拟合
贝叶斯方法	基于贝叶斯定理的一类算法	朴素贝叶斯算法, 贝叶斯(信念)网络, 贝叶斯神经网络	一定程度上反应变量之间的因果关系; 能够估预测结果的不确定性; 能够在小样本情况下取得很好的效果
基于核的算法	把输入的原始空间的特征映射到高维甚至无穷维空间中,在高维空间学习得到对应的线性模型的一类算法	支持向量机, 相关向量机, 核岭回归	适合小样本数据; 可以解决高维问题; 具有较好的鲁棒性和泛化性; 对参数比较和数据噪声敏感; 难以用于大规模样本训练
遗传算法	基于达尔文进化论的全局寻优算法,通过编码,交叉,变异,根据适应度函数搜索寻优,找到最优解的算法	—	适用于复杂和困难场景,如带有大量噪声和无关数据、事物不断更新、目标无法精确定义等; 具有较好的鲁棒性
集成算法	集成多个弱学习器得到的强学习器	随机森林, 梯度提升决策树, 堆叠泛化	灵活可变,性能优越; 减少了过拟合
人工神经网络	模拟生物神经网络建立的、由大量处理单元互联组成的非线性运算模型	反向传播神经网络, 广义回归神经网络	学习能力优越; “黑匣子”特性与不可解释性
深度学习	含有多层隐含层的神经网络模型	深度神经网络, 深度信念网络, 卷积网络, 长短期记忆神经网络, 图神经网络	强大的学习能力; 能够处理多种多样的复杂问题; 依赖于大量的数据作为训练; 计算成本相对高昂; “黑匣子”特性与不可解释性

## 2.1 回归算法

定量遥感领域一种常用的较为简单的回归算法是线性回归 (Linear Regression)。线性回归通过构建卫星观测和目标参量之间的线性关系,求解损失函数最小时的参数 (Uyanık 和 Güler, 2013),实现由卫星观测到目标参量之间的反演。因建模快速简单的特点,线性回归在早期被广泛应用于各个领域 (Wang 和 Christopher, 2003; Hyypä 等, 2000; Tai 等, 2010)。例如 Wang 和 Christopher

(2003) 构建 MODIS AOD 产品与  $PM_{2.5}$  之间的线性关系,用于  $PM_{2.5}$  浓度的反演; Tai 等 (2010) 在此基础上进一步引入气象要素,构建了多元线性回归模型; Hyypä 等 (2000) 同样利用多元线性回归模型,从 Landsat TM, Spot PAN 和 XS 等卫星数据中反演得到了林分属性。除此之外,标准线性回归也在被不断的优化和改进,衍生出了诸多特点各异的线性回归算法。例如考虑到目标参量和卫星观测数据之间的关系会随着时间和空间而改

变, 学者们进一步引入考虑了时空异质性的线性混合效应模型 (Lee 等, 2011) 和时空地理加权回归模型 (Guo 等, 2017; Qin 等, 2017) 以提升定量反演精度。

但线性回归方法不太适合处理具有非线性关系的数据, 多项式回归 (Polynomial Regression) (Edwards 和 Parry, 1993) 则解决了这个问题。多项式回归的最佳拟合线不是一条直线, 而是一条符合数据点的曲线, 且一些自变量的指数是大于1的。相较于线性回归, 多项式回归总体上更灵活, 可以模拟一些复杂的关系。Cannizzaro 和 Carder (2006) 利用三次多项式函数拟合不同波段反射率比值与叶绿素 a 浓度之间的关系, 实现叶绿素 a 浓度的遥感估算; Sawaya 等 (2003) 利用二次多项式拟合不透水表面面积百分比和 IKONOS NDVI 之间的关系, 实现了从 NDVI 到不透水面的反演; Wang 等 (2013) 同样利用三次多项式, 从窄波段植被水分和干物质指数比值反演得到了燃料水分含量。

在进行回归预测时, 如果变量较多, 往往要进行变量选择, 得到最优的模型输入, 因此, 有学者提出了逐步式回归 (Stepwise Regression) 的策略。其基本思想是将变量逐个引入, 引入条件是其偏回归平方和经检验后是显著的。同时, 每引入一个新变量后, 对已入选回归模型的老变量逐个进行检验, 将经检验认为不显著的变量删除, 重复此过程直至不能引入新变量, 这时回归模型中解释变量集为最优。因其对输入变量的选择和控制, 逐步式回归往往比非逐步式的方法可以取得更优的结果, Chen 等 (2018) 利用 Landsat TM 数据对灌木生物量估算的实验也证明了此观点。然而, Grossman 等 (1996) 在其用叶片反射率数据反演叶片生化信息的实验中发现, 虽然逐步式回归可以得到最优的解释变量集, 但是其选择出的波段与真实的吸收光谱并不相符, 而且具有较大的随机性, 选出的最佳波段会随着数据集、因变量类型的变化而变化。因此, 作者呼吁, 在鲜叶反射光谱问题上, 应当谨慎的使用逐步式回归。

多元自适应回归样条 MARS (Multivariate Adaptive Regression Splines) 是逐步式回归的进一步推广, 由美国的统计学家 Friedman (1991) 年提出。MARS 是一种基于分段策略的非线性非参数回归方法,

该方法以样条函数的张量积作为基函数, 分为前向过程、后向剪枝过程两个步骤。其优势在于能够处理数据量大、维度高的数据, 而且计算快捷、模型精确, 可解释性强。在 Kuter 等 (2018) 利用 MODIS 数据反演积雪覆盖比的实验中, MARS 取得了和神经网络算法相媲美的精度, 与此同时, MARS 的计算效率远高于神经网络算法。Huang 等 (2020) 利用 MARS 进行河流组分浓度等反演, 并基于 MARS 算法良好的可解释性, 详细讨论了自变量和因变量之间的交互作用。Islam 等 (2015) 利用 MARS 和微波遥感数据进行降雨率的反演, 并认为 MARS 具有高效、灵活的特点, 将十分适用于全球降雨数据的反演。在地上生物量定量反演的实验中, 作者利用高光谱数据及其他辅助变量作为输入, 结果表明, MARS 在高维数据下也可以取得较好的反演结果 (Filippi 等, 2014)。

## 2.2 正则化方法

正则化方法一般指向模型中引入附加约束以防止过拟合和提升模型性能的一类方法 (Bickel 等, 2006)。例如, 为了解决标准线性和多项式回归在特征变量共线性的情况下会失败的问题, 发展出了岭回归 (Ridge Regression) (Hoerl 和 Kennard, 1970) 和 Lasso 回归 (Least Absolute Shrinkage and Selection Operator) (Kukreja 等, 2006), 以及弹性网络 (Elastic Net) (Zou 和 Hastie, 2005)。它们都是在回归优化函数中增加一个偏置项, 以减少共线性的影响, 从而减少模型方差。区别在于岭回归使用 L2 正则化, Lasso 回归使用 L1 正则化, 而弹性网络 ElasticNet 同时使用 L1 和 L2 正则化。这些正则化方法在输入变量具有高维度和多重共线性的情况下也能有良好的效果 Kang 等 (2018) 利用这一特性, 通过引入岭回归, 解决了土壤湿度站点之间的共线性问题, 实现了对稀疏土壤湿度观测的空间升尺度。除此之外, 正则化作为一种灵活的约束方法, 可以以不同的形式与多种不同的模型相结合。有许多学者根据待解决问题的特点, 将一些先验知识作为正则化项加入到机器学习算法中, 以提升模型性能。例如, Atzberger (2004) 和 Combal 等 (2003) 将辐射信息作为正则化所需的先验信息, 提升了基于人工神经网络的冠层参数反演模型的性能。Hang 等 (2017) 提出图正则化非线性岭回归算法 (KGRR), 并应用于

高光谱分类和气溶胶反演任务。结果表明,相比于普通的岭回归算法,支持向量机,以及神经网络算法,KGRR在分类和回归任务中均具有更优的表现。未来,这种根据特定问题改进的机器学习算法,以及多类算法融合的模式,将成为学者们重点关注的方向之一。

### 2.3 基于实例的算法

不同于大多数机器学习算法都有围绕目标函数建立起来的明确的一般化描述,基于实例的算法(Instance-based learning)直接通过存储的数据集进行分类或回归来得到学习结果。代表性算法包括K邻近算法(KNN),自组织映射算法SOM(Self-Organizing Map),局部加权回归等。定量遥感领域的应用包括利用K邻近回归(KNR)进行叶面积指数(Wei等,2017)、生物量(Antropov等,2017)的反演;利用SOM算法实现二次浮游植物色素(El Hourany等,2019)、叶绿素a垂直分布(Charantonis等,2015)的估算,以及总悬浮颗粒物浓度产品的评估(Gao等,2019)等。但基于实例的算法内存使用往往较高、计算成本高,无法用于高维特征空间,因此应用也较为受限。

### 2.4 决策树学习

决策树算法根据数据的属性采用树状结构建立模型,常见的算法包括:ID3(Iterative Dichotomiser 3),CART(Classification And Regression Tree),以及C4.5。由于单棵决策树的学习能力有限(不如决策树的集成算法如随机森林),因此在定量遥感领域的应用较为有限,主要包括用于植被类型、土壤类型、积雪判别等分类问题(Elnaggar和Noller,2009;Ghulam等,2014;Grody和Basist,1996),以及少量的回归反演任务中(Caicedo等,2014;Jiang等,2021)。

### 2.5 贝叶斯方法

贝叶斯方法算法是基于贝叶斯定理的一类算法,常见的包括:朴素贝叶斯算法,平均单依赖估计AODE(Averaged One-Dependence Estimators),以及贝叶斯(信念)网络BBN(Bayesian Belief Network)。前两者主要用于分类问题,因此在定量遥感反演中应用较少。而贝叶斯网络则应用较多,尤其是在植被参数定量反演遥感问题上(Qu等,2008;Xu等,2019;Zhang等,2012),其基本原

理是利用叶面积指数、叶绿素浓度、光谱反射率、卫星观测角等参数,构建一个有向无环图,利用该概率图模型,结合贝叶斯理论,完成对植被参数的定量反演。由于贝叶斯模型可以在一定程度上反应变量之间的因果关系,并根据已有的知识做出推断,近年来,其与深度学习模型的结合也成为了一大研究热点,贝叶斯神经网络BNN(Bayesian Neural Network)也因此诞生。其核心思想是将神经网络中的权重看作某种分布,而非固定值。由于其预测值也是概率分布,因此,BNN可以度量预测结果的不确定性。Ma等(2021)利用BNN,结合MODIS数据,气候数据和土壤属性数据,实现了对玉米产量的预测,并对预测值的不确定性进行了评估。BNN不仅能够评估预测结果的不确定性,也能够在小样本情况下取得很好的效果(Ma等,2021),然而,目前基于BNN进行参数定量反演的工作还相对较少,值得进步一步探索和挖掘。

### 2.6 基于核的算法

基于核的算法把输入的原始空间的特征映射到高维甚至无穷维空间中,并在高维空间学习得到对应的线性模型,使模型具有极强的非线性表达能力。其中最著名的当属支持向量机SVM(Support Vector Machine)。SVM最初主要被用于分类任务,后来Vapnik等(1996)发展出了用于回归的SVM算法,称为支持向量回归(SVR)。SVM算法适合小数量样本数据;算法设计很好的避免了维度灾难,可以解决高维问题;不但简单,而且具有较好的鲁棒性和泛化性。因此,20世纪以来,深入地球和遥感科学的各个领域,得到颇为广泛的应用。其他基于核的算法也包括相关向量机(RVM),核岭回归(KRR)等,也已被应用于陆地、海洋和大气等多个领域(Camps-Valls和Bruzzone,2009;Verrelst等,2015)。Camps-Valls和Bruzzone(2009)在其专著中对基于核的算法在遥感领域中的应用做出了系统的总结,认为基于核的算法已经成为了遥感应应用领域的标准方法之一。Mountrakis等(2011)在其综述论文中对SVM算法在遥感中的应用,以及SVM算法的优缺点做出了总结。作者指出,虽然SVM算法优势众多,但其依然存在参数敏感、噪声敏感,难以用于大规模样本训练的问题。因此,在地球大数据时代(李德仁等,2014a),面

对海量的待处理数据, SVM算法也存在一定的局限性, 需要进一步的改进和提升 (Foody, 2008)。

## 2.7 遗传算法

遗传算法是一种基于达尔文进化论的全局寻优算法, 最早由 Holland 在 1969 年提出, 后经多位学者整理形成 (De Jong, 1975; Goldberg 和 Holland 1988; Holland, 1975)。遗传算法首先通过编码将待求解的问题表示成遗传空间的染色体或者个体, 令个体之间进行交叉和变异; 同时根据所求问题的目标函数来设计适应度函数, 使得算法不断对编码后的参数进行搜索寻优, 直到找到最优解或达到最大进化代数。遗传算法适用于非常复杂和困难的环境, 比如, 带有大量噪声和无关数据、事物不断更新、问题目标不能明显和精确地定义等; 且具有较好的鲁棒性。在定量遥感领域, 遗传算法也被用于土地利用分类 (Tseng 等, 2008) 和冰云属性反演 (Liu 等, 2021) 等问题。除此之外, 作为一种全局优化算法, 遗传算法也能很好的与其他算法相融合。例如, 与偏最小二乘算法耦合 (GA-PLS), 从遥感反射率数据反演植被等效水厚度和鲜叶片燃料水分含量 (Li 等, 2008, 2007); 与极值梯度提升树耦合 (XGBR-GA) (Pham 等, 2020), 利用多源遥感数据反演得到地上生物量, 并且取得了比梯度提升树、支持向量机、随机森林更好的精度; 以及与神经网络算法耦合 (GA-BP), 协同主被动微波遥感, 反演土壤水分。作者认为, 遗传算法很好地弥补了神经网络容易陷入局部最小的缺点, 且使神经网络具有较快的收敛性、全局优化的能力和较强的学习能力 (余凡等, 2012)。

## 2.8 集成算法

上述算法主要都是单个学习器, 他们在性能提升上往往存在一定的瓶颈, 也可能容易出现过拟合或欠拟合的问题; 因此, 集成多个学习器来提高模型性能的方法应运而生, 这就是集成学习算法 (Ensemble Learning)。集成算法根据集成方式的不同一般分为 3 类: (1) Bagging。以并行的方式把弱学习器的结果组织起来, 通过投票或取平均等折中的方式取得最终的结果, 代表性算法为随机森林 (Random Forest)。(2) Boosting。以串联的方式组织模型, 根据旧模型中的错误来训练

新模型, 层层改进, 代表性算法为 AdaBoost、梯度提升决策树 GBDT (Gradient Boosting Decision Tree)。(3) Stacking。使用上一阶段的预测结果作为下一层预测的输入特征, 重新训练一个新的学习器, 输出最终的结果。代表性算法为堆叠泛化 (Stacked Generalization)。集成算法的泛化能力和预测性能通常比单个学习者更强大, 因此应用十分广泛。从陆地遥感, 到海洋遥感, 到大气遥感, 无论是分类问题, 还是定量反演问题, 集成学习等法, 都得到了广泛的应用。(Healey 等, 2018; Lee 等, 2019; Wang 等, 2019)。集成算法优越的性能使得它成为了机器学习领域重要的发展方向之一, 其在定量遥感领域的应用潜力也有进一步挖掘的价值, 值得重点关注。

## 2.9 人工神经网络

作为过去十年中最具影响力的技术之一, 人工神经网络收到了空前广泛的关注。人工智能先驱的最初愿景是复制人类大脑的功能, 人工神经网络便是模拟生物神经网络的产物。它由输入层, 隐藏层和输出层组成, 其中输入层从外部接收数据, 一个或多个隐藏层处理数据, 输出层提供一个或多个神经元输出结果。人工神经网络是机器学习的一个庞大的分支, 有几百种不同的算法, 深度学习就属于其中之一, 将在下节具体讨论, 因此, 本节的神经网络算法主要指非深度学习的浅层网络。神经网络算法十分灵活, 可以快速调整, 适应各种不同类型的新问题, 因此, 同集成算法一样, 应用十分之广泛。其在定量遥感领域的应用, 可以追溯到 20 世纪 90 年代, 最初主要被用于遥感影像分类问题 (Parikh 等, 1990), 不久后也被引入定量反演问题中 (Davis 等, 1993)。其中, 最为常用的是具有多层感知器结构的反向传播神经网络 BPNN (Back Propagation Neural Network) (Dong 等, 2020; Tedesco 等, 2004)。除此之外, 自组织映射 (Chang 和 Islam, 2000)、广义回归神经网络 (Li 等, 2017) 等也有一定的应用。神经网络具有高度的灵活性, 近几年, 根据特定的问题而设计特定的神经网络, 受到了诸多学者的关注。Li 等 (2020a) 将时空异质性融入广义回归神经网络, 提出了时空加权神经网络, 充分考虑了  $PM_{2.5}$  浓度和 AOD 关系的时空差异, 提升了传统广义回归神经网络的效果。Wu 等 (2021) 将这种时

空加权策略和时空邻近神经网络相融合,提出的时空神经网络加权回归算法,同样取得了比传统算法更优异的表现。这种考虑待解问题特性的而设计的神经网络,将是一个重要的发展方向之一。

### 2.10 深度学习

深度学习算法是对人工神经网络的进一步发展。近年来,随着计算能力的不断提升,深度学习迎来了飞速的发展,网络结构日趋复杂。深度学习在定量遥感领域的起步不久,却已经取得了突出的成就,受到了广泛的关注。深度神经网络DNN(Deep Neural Network)(Su等,2020),深度信念网络DBN(Deep Belief Networks)(Shen等,2018),卷积网络CNN(Convolutional Neural Network)(Malmgren-Hansen等,2019),长短期记忆神经网络LSTM(Long Short-Term Memory)(Ma等,2019)、图神经网络GNN(Graph Neural Network)均得到了应用和发展。Yuan等(2020)对遥感领域常用的深度学习模型结构,深度学习在环境遥感中的具体应用,以及目前面临的挑战和可能的发展方向,做出了系统的总结:相比于深度学习算法的发展速度,深度学习在定量遥感领域的应用还有待深入,迁移学习、生成对抗学习、小样本学习、耦合物理规律的深度学习等问题仍有待探索。

## 3 数据驱动和模型驱动的结合

基于数据及归纳的人工智能方法(机器学习,Machine Learning)可以从不断增长的地理空间数据流中捕捉信息、提取模式(Reichstein等,2019),有很强的数据适应性,但同时,机器学习缺乏理论支持、可解释性弱等“黑匣子”问题亟待解决。因此,传统的物理模型的发展也不容忽视。不同于机器学习,理论模型基于知识及演绎,包含丰富的底层系统知识,其结果可解释性强(Boukabara等,2021)。在地球科学系统中,两者虽驱动力不同,却具有相同目的,即通过数据分析或数学建模来预测行为(Dueben和Bauer,2018)。故在大数据背景下,数据驱动方法和模型驱动方法的优势结合成为定量遥感各研究领域的热点。

基于不同的主体,两者结合的相关工作可大致分为两组:机器学习辅助的模型模拟和模型辅助的机器学习。前者利用机器学习优化理论模型,进行集成建模;后者将先验知识整合到机器学习中,

可描述为知情机器学习(von Rueden等,2020)。

### 3.1 机器学习辅助的模型模拟

模型模拟是一种自上而下的模型驱动方法,具有坚实的理论框架。从系统的角度出发,机器学习可从原始系统、子模型、参数3个层次对其进行优化(图3)。如图3所示,优化系统为利用机器学习方法纠正数据同化过程中可能存在的系统误差;优化子模型为训练XGBoost模型替代大气化学传输模型中的化学积分过程(Keller和Evans,2019),以提高效率;优化参数为基于神经网络模拟气溶胶物理参数 $VE_r$ 和相关变量的函数关系,从而优化PMRS大气物理模型(Zhang和Li,2015)。

#### 3.1.1 优化系统:校正系统偏差

假设没有观测偏差,模型模拟与观测值的偏差可归因于不完善的知识而产生的系统误差。机器学习能有效识别、提取并可视化模型错误的模式,从而纠正相应初始动态变量的隐式系统误差(Reichstein等,2019)。近年,Pathak等(2018)探索了混合“数据辅助”动力学模型在预测方案中校正偏差的可能性。进一步,机器学习辅助优化有助于填补系统知识空白、探索物理方程进而改进理论模型。

数据同化(Data Assimilation)使得观测与模型协同发展,是地球科学系统中的一个重大领域(Eyre等,2020)。它以物理模型为基础,通过融合多源观测数据进行系统状态的估计、观测—模型的不确定度评估,从而提高系统的可预测性(Li等,2020b)。Geer(2021)分析了贝叶斯框架下数据同化和机器学习的统一性,机器学习的融入为搜索数据同化中统计度量的最佳拟合提供了可行方案。例如,Irrgang等(2020)提出了一种基于机器学习的大气强迫不确定性估计方法,其结果作为数据同化中误差协方差信息的来源;Cintra和de Campos Velho(2014)利用机器学习模拟了Kalman滤波方案从而嵌入大气环流模式;Ruckstuhl等(2021)训练卷积神经网络来维持数据同化中的质量守恒,从而增强了遥感数据、地面数据与模型模拟结果的一致性。

#### 3.1.2 优化子模型:代理建模

模型模拟的大框架往往由许多子模型构成,如果子模型是半经验性质的,其函数形式理论基

础较弱(例如,生物过程),代理建模可在保持能量守恒定律等物理规律的前提下,实现子模型和全模型的误差分离,从而优化建模。另外,偏微分方程求解的不确定性问题也是模型发展存在的问题之一(Zanna等,2019)。此思路已被用于优化海水运动场,De Bezenac等(2018)建立了卷积-反卷积神经网络以学习海水的运动场,然后使用物理模型进一步处理以更新热含量和温度。Mo等(2019)利用深卷积编解码网络替代传统的蒙特卡罗(Monte Carlo)方法量化动态多相流问题的不确定性,提升了地下水模型的预测效率(如压

力和饱和度)。在大气科学领域,Keller和Evans(2019)将机器学习如随机森林方法引入大气化学传输模式,通过替代化学积分器这一子过程简化化学微分方程的计算,提升计算速度的同时,输出的大气污染物浓度与原模式精度相当如图3中优化环节所示。Rasp等(2018)将神经网络用于气候模型SPCAM(Super-Parameterized Community Atmosphere Model v3.0)中的大气亚格网过程的近似,具有快速准确的优势,显示了数据驱动下气候模式开发的潜力。

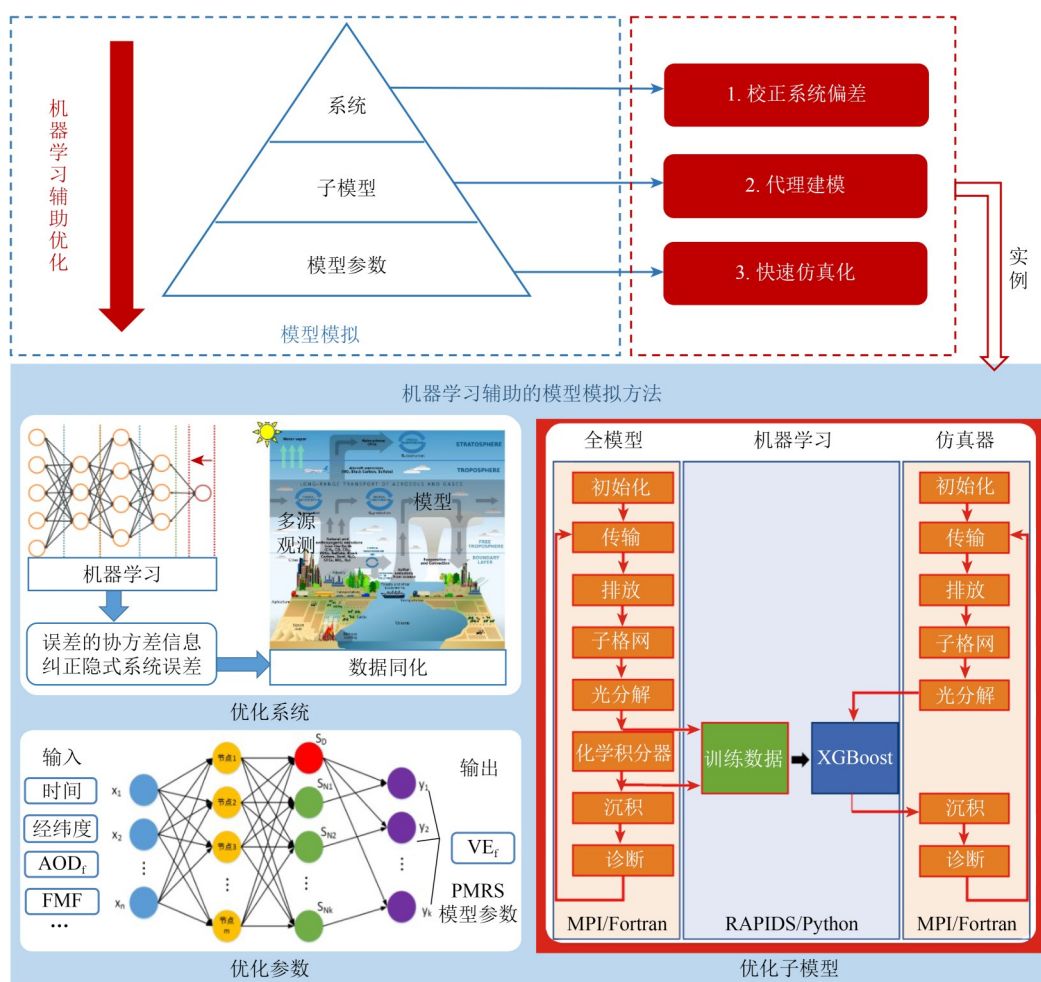


图3 机器学习辅助的模型模拟:3种优化思路

Fig. 3 Three optimization ideas for model simulation assisted by machine learning

### 3.1.3 优化参数:快速仿真化

机器学习可提供误差相对较小的模型参数的函数近似(Veerman等,2021),实现快速仿真。针对气候模式难以表征的湿对流相关参数,Gentine等(2018)提出了一种新的基于机器学习的对流参

数化方法,其从明确的模拟中学习对流的显著特征,计算效率显著高于超参数化;基于准地转海洋模式(Quasi-Geostrophic Ocean Model),Bolton和Zanna(2019)成功利用神经网络表示涡流动量强迫的时空变异性,对动量的涡动参数化和更普遍



的亚网格参数化有重要意义；目前，大气辐射参数和行星边界层的机器学习模拟也应用广泛，研究表明短波辐射计算因机器学习的参数仿真加快了约60倍（Boukabara, 2020）。Sawada (2020) 结合微波亮温数据和MCMC (Markov Chain Monte Carlo) 算法，使用统计机器学习的代理建模技术加速了陆地表面模型 (Land Surface Model) 中土地湿度和植被含水量相关的4个参数优化和不确定性评估。

此外，在实现快速仿真的同时，Yuval 和 O'Gorman (2020) 通过优化高分辨率模型训练数据，探索了机器学习应用于气候建模中亚网格过程的稳定参数化的可能性。对于大气领域的物理模型来说，代表气溶胶光学性质的相关物理参数对准确反演大气污染物（如 $PM_{2.5}$ ）至关重要。Zhang 和 Li (2015) 提出一种不依赖于地面站点数据的颗粒物反演模型 PMRS (Particulate Matter Remote Sensing

method)，并将关键物理参数—柱状体积消光比  $VE_r$  (columnar volume-to extinction ratio of fine particulates) 拟合为有关细粒子分数的二次多项式。 $VE_r$  包含复杂的粒子微物理性质，且与光学参数紧密联系，基于此，神经网络可建立  $VE_r$  与多个相关变量之间的模型，例如遥感卫星提供的细粒子分数 FMF (Fine Mode Fraction)、AOD 等。图3中优化参数环节展现了具体思路，通过捕捉上述变量、时空变量和  $VE_r$  之间的复杂关系，优化该物理参数的表达，从而进一步提升  $PM_{2.5}$  反演的精度。

### 3.2 模型辅助的机器学习

机器学习以数据作为驱动力，“自下而上”地刻画地球科学系统中各变量之间复杂的非线性关系。将物理模型有关的先验知识引入其中，有利于提高它的合理性及可解释性。从机器学习训练的角度出发，模式可从数据、过程、结果等3个方面对其进行增强（图4）。

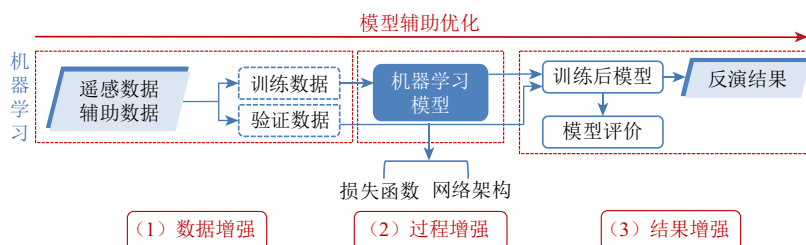


图4 模型辅助的机器学习的3种优化思路

Fig. 4 Three optimization ideas for model-assisted machine learning

#### 3.2.1 数据增强

现实观测数据（例如遥感数据、地面数据等）通常具有非平稳性，而模型却没有直接的观测限制（van Donkelaar 等, 2015）。将模型模拟数据作为机器学习的附加数据来源，数据集得以扩展，物理性知识融入训练过程，数据质量得以增强。

在大气污染物浓度反演过程中，化学传输模型 CTM (Chemical Transport Model) 可提供关于历史污染物浓度额外有价值的信息。例如，将 CTM 的模拟  $PM_{2.5}$  值或  $PM_{2.5}$  组分信息纳入总数据集，与卫星 AOD 数据、气象变量等一同建模，其为  $PM_{2.5}$  的时空变化提供了先验知识（LYU 等, 2019）。为结合观测和多模型输出，Chang 等 (2019) 探索了一种新的方法 (M3Fusion v1) 以改进全球表面臭氧分布估计，进一步可利用机器学习优化。在水文领域，Yang 等 (2020) 为填充稀疏观测开发了

物理过程和神经网络组合的水文模型，其在有限的观测数据下能模拟大流域的日流量，并探究了合成数据驱动的机器学习模型的性能优越性；Yin 等 (2021) 融入地表水流系统输出，用贝叶斯机器学习集成方法较好地预测了地下水蓄水量变化。另外，Berger 等 (2020) 通过耦合叶片光学特性模型 PROSPECT-PRO 与冠层反射率模型 4SAIL 生成了一个训练数据库，基于概率机器学习回归方法，提出了一种估计农作物氮含量的混合反演方法。目前高光谱及多光谱图像是遥感监测氮含量的主要数据来源，所建立的模型具有较强的泛化能力、较大的时空覆盖范围（陈永喆 等, 2017）。结合多源数据优势，耦合定量遥感和物理模型方法进行反演，也将成为未来研究的一大热点。

#### 3.2.2 过程增强

目前，生成对抗网络 GAN (Generative Adversarial

Networks) 得到迅速发展以实现将模拟集成到学习算法中的功能, 例如, PI-GANs (Wu等, 2020) 确保了生成样本的气候协方差与训练数据集的气候协方差相匹配。理论上, 模型能引导机器学习的训练过程, 给予算法相应的物理约束, 基于此在不同领域也开发了不同的混合机器学习模型。具体方式主要分为以下两种。

(1) 较为流行的方法是将物理约束作为损失函数的附加项, 增强学习结果的物理一致性。Beucler等(2020)使用温度—密度—深度的物理转换关系设计神经网络损失函数的一部分, 基于湖泊模型模拟构建了物理引导的神经网络(PGNN)以反演水温。在此基础上, Jia等(2018)考虑LSTM模型的时间建模结构, 增加了对预测水温随时间变化的约束, 使训练过程遵循能量守恒。Zhao等(2019)将基于能量平衡、扩散理论的二次Penman-Monteith方程嵌入损失函数以反演蒸发模型中的潜在热通量, 实验证明该网络具有更高的泛化能力。Beucler等(2020)以约束矩阵的形式反映对流模型中需要遵循的守恒规律, 设计损失函数中的附加项并进行加权组合, 相比于原始网络, 对流增湿变量的预测误差显著降低。

(2) 通过更改网络架构, 将模型中的物理规律整合入训练过程, 从而形成约束层。Jiang等(2020a)构造了一个特殊循环神经层来包装物理模型(时间动态地球科学模型), 作为深度学习架构中一部分进行训练, 输出动态响应(如每日的径流量), 中间变量例如积雪蓄水量也得到较高精度的模拟。同时, Beucler等(2020)探究了ACnet网络以预测对流增湿和增热变量, 平衡了模型精度和物理一致性, 网络把质量、能量、辐射守恒定律整合于约束层以输出额外的变量值, 计算残差从而协同优化网络。为预测云过程(即对流和辐射)对气候影响的多维向量。Beucler等(2019)在构建网络时纳入4种守恒定律, 同时提出设计损失函数和更改网络架构两种方法, 前者合理选择物理附加项的加权值, 后者通过求解约束方程组计算输出向量的4个分量, 结果表明物理约束信息可以显著提高模型泛化能力。此优化思路同样可应用于其他领域, 通过变换与各反演变量相关的物理模型的表达形式, 将其合理嵌入神经网络以实现反演, 理论上遥感模型的精度及普适性将得到进一步提升。

上述模型辅助机器学习的过程增强使得整个任务序列是通过统一的人工智能架构完成, 从结构上弥补了机器学习的物理不一致问题。此外, 基于深度学习的灵活性和易于扩展性, 常规神经网络层也可置于物理约束层之上以提供输入, 例如云的参数化(Rasp等, 2018), Jiang等(2020b)对网络施加一组偏微分方程约束, 搭建了一个时空超分辨率框架以从低分辨率输入中生成高分辨率时空解, 此研究对于定量遥感也具有一定意义。

### 3.2.3 结果增强

神经网络中的性能度量可用两种方法来表示, 一种是损失函数, 另一种是辅助度量, 后者通过定义和跟踪一个或多个额外的期望性能标准(不必可微), 在每次网络训练后对其进行评估, 以确保立即检测到不可接受的权衡(Ebert-Uphoff和Hilburn, 2020)。为了充分挖掘降水的潜在降尺度解, Chen等(2020)设计了一个学习降水图的隐式物理运动估计框架(RainNet)并将8个基于数据集物理性质的度量标准纳入评价标准。从这个角度看, 模型约束可作为机器学习的结果增强手段。

其次, 可从网络解空间的角度出发进行增强。Kraft等(2020)提出了一个端到端的全球混合水文模型, 其目的为反演总蓄水量、径流量、蒸散量以及雪水当量。模型将概念性水平衡模型机理融入动态神经网络的解空间, 使相关变量保持可解释性。从深度学习的角度来看, 混合方法可看作是神经网络的正则化, 将解空间限制为物理上合理的结果。同时, 复杂物理模型的大范围动力学模拟(如云分辨率模型)对神经网络结果的测试也奠定了坚实的基础(Pal等, 2019)。

### 3.3 两者融合的问题与挑战

任何直接从数据中学习的方式都必须应对地球科学发展的挑战(Geer, 2021), 其中包括以定量的方式表示不确定性、建立观测数据与地学过程的联系等。在大数据时代, 二者融合同样面临类似问题。

(1) 物理模型内部机理的复杂性。优化模型中的某一步骤或参数, 是近年来机器学习嵌入物理模型中的主流方法, 故明确各领域有关参数意义至关重要。解决此问题需促进地理学、生物学、

人工智能学等众多领域交叉合作, 探寻与已知或未知参数相关的变量, 赋予数据可解释性从而以简代繁, 使机器学习有针对性地提升精度。

(2) 联合模型的不确定性。将物理意义与大数据驱动思路有效结合, 如何减少联合模型的不确定性是需要攻克的一大难关。例如, 在带有物理约束条件的深度学习训练中, 若地面测量数据与物理模型输出值同为训练真值, 则需合理分配二者占比, 降低网络学习中的不确定性, 从而使其更好地协同优化网络。另外, 定量指标是评价模型结果的重要标准, 机器学习注重精度, 物理模型强调地学意义, 在训练模型过程如何均衡考虑两类指标以优化建模尤为关键。

(3) 联合模型的泛化能力。由于样本的限制, 机器学习方法在小区域内表现良好, 泛化能力因数据性质及训练方法而有所差别, 当适用于普遍数据的物理模型嵌入后, 模型是否能在大范围内保持良好的预测能力, 仍有待验证。

(4) 联合模型的可迁移性。不同领域研究基于的物理模型、观测数据类型不一致, 应用的人工智能方法也大相径庭, 当联合模型的优良架构(如GAN)出现, 是否可以在理论相近的子领域进行模型迁移, 例如云辐射、气溶胶、颗粒物等大气相关研究方向, 此思路则需要多个领域共同探究并深入合作。

(5) 小样本情况下的联合建模问题。物理模型中的相关参数往往表征了复杂的物理特性, 且与多种性质相关, 故某些物理模型的中间参数难以通过观测获得, 即训练真值数量较少。当样本稀疏时, 需要通过合适的机器学习方法来表达、模拟及验证参数, 此外, 不拘泥于该替代建模步骤的样本验证, 也可对后续步骤的精度提升效果做出定量评价, 判断机器学习训练的模型是否可靠。

物理模型包含丰富的可解释性知识, 多源数据的融入有助于改善理论、优化系统, 机器学习的辅助有利于其效率、精度的提升。虽然放弃物理模型知识并从零开始使用机器学习很有诱惑力(Sønderby等, 2020), 但考虑模型理论并将其嵌入学习过程会从先前的知识中获益, 从而优化反演结果。实现数据驱动和模型驱动方法的协调发展, 使两者相互促进、相互融合, 也是深入研究地球科学系统各领域问题的大势所趋。

## 4 结 语

定量遥感可以看作是将数据转化为信息的过程。模型驱动的定量遥感通过现有认知, 如物理规律、生化性质等, 从数据中提炼信息, 形成新的认知。是一个旧认知产生新认知, 旧模型驱动新模型, 从而不断完善对地球系统的认知和建模的过程。数据驱动的定量遥感则通过对数据进行“学习”来得到目标信息, 它不关心现有的认知和理论框架, 而关心如何“学习”。从简单的一元线性回归, 到集成学习、深度学习; 从单一场景、单一要素的物理建模, 到考虑多场景, 多要素, 数据驱动和模型驱动的方法都经历了长足的发展, 在不断进步和完善。然而, 随着对两类模型研究的深入, 他们的各自的局限性也日渐凸显, 数据驱动的方法忽视内在机理, 弱化因果关系; 模型驱动的方法不确定性大, 成本高昂。基于此, 耦合物理规律和机器学习, 构建模型和数据共同驱动的反演框架, 成为了定量遥感领域新的研究热点。现有研究或用机器学习优化物理模型, 或用物理规律增强机器学习的预测能力, 均取得了一定的成效。然而, 当前的耦合仍然为浅层的耦合, 对复杂的物理模型内部机理挖掘尚浅, 模型效果提升较为有限。除此之外, 联合模型的不确定性、泛化能力、可迁移性、小样本情况下的适用性仍有待探讨。未来, 探索深度耦合的新思路, 新方法, 将是定量遥感发展的重要方向之一。

## 参考文献(References)

- Antropov O, Rauste Y, Häme T and Praks J. 2017. Polarimetric ALOS PALSAR time series in mapping biomass of boreal forests. *Remote Sensing*, 9(10): 999 [DOI: 10.3390/RS9100999]
- Atzberger C. 2004. Object-based retrieval of biophysical canopy variables using artificial neural nets and radiative transfer models. *Remote Sensing of Environment*, 93(1/2): 53-67 [DOI: 10.1016/j.rse.2004.06.016]
- Berger K, Verrelst J, Féret J, Hank T, Woche M, Mauser W and Camps-Valls G. 2020. Retrieval of aboveground crop nitrogen content with a hybrid machine learning method. *International Journal of Applied Earth Observation and Geoinformation*, 92: 102174 [DOI: 10.1016/j.jag.2020.102174]
- Beucler T, Pritchard M, Gentine P and Rasp S. 2020. Towards physically-consistent, data-driven models of convection//IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Sympo-

- sium. Waikoloa, HI, USA: IEEE: 3987-3990 [DOI: 10.1109/IGARSS39084.2020.9324569]
- Beucler T, Rasp S, Pritchard M and Gentine P. 2019. Achieving conservation of energy in neural network emulators for climate modeling. arXiv: 1906.06622
- Bickel P J, Li B, Tsybakov A B, Van De Geer S A, Yu B, Valdés T, Rivero C, Fan J Q and Van Der Vaart A. 2006. Regularization in statistics. *Test*, 15(2): 271-344 [DOI: 10.1007/BF02607055]
- Bolton T and Zanna L. 2019. Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, 11(1): 376-399 [DOI: 10.1029/2018MS001472]
- Boukabara S A, Krasnopolsky V, Penny S G, Stewart J Q, McGovern A, Hall D, Ten Hoeve J E, Hickey J, Allen Huang H L, Williams J K, Ide K, Tissot P, Haupt S E, Casey K S, Oza N, Geer A J, Maddy E S and Hoffman R N. 2021. Outlook for exploiting artificial intelligence in the earth and environmental sciences. *Bulletin of the American Meteorological Society*, 102(5): E1016-E1032 [DOI: 10.1175/BAMS-D-20-0031.1]
- Boukabara S A, Krasnopolsky V, Stewart J Q, Maddy E, Shahroudi N, and Hoffman R N. 2020. Realizing the Benefits of AI across the Numerical Weather Prediction Value Chain, *Bulletin of the American Meteorological Society*, 101(1), 29-33 (Caicedo J P R, Verrelst J, Muñoz-Marí J, Moreno J and Camps-Valls G. 2014. Toward a semiautomatic machine learning retrieval of biophysical parameters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(4): 1249-1259 [DOI: 10.1109/JSTARS.2014.2298752]
- Camps-Valls G and Bruzzone L. 2009. *Kernel Methods for Remote Sensing Data Analysis*. Chichester: John Wiley and Sons
- Cannizzaro J P and Carder K L. 2006. Estimating chlorophyll a concentrations from remote-sensing reflectance in optically shallow waters. *Remote Sensing of Environment*, 101(1): 13-24 [DOI: 10.1016/j.rse.2005.12.002]
- Chang D H and Islam S. 2000. Estimation of soil physical properties using remote sensing and artificial neural network. *Remote Sensing of Environment*, 74(3): 534-544 [DOI: 10.1016/S0034-4257(00)00144-9]
- Chang K L, Cooper O R, West J J, Serre M L, Schultz M G, Lin M Y, Maréchal V, Josse B, Deushi M, Sudo K, Liu J H and Keller C A. 2019. A new method (M<sup>3</sup> Fusion v1) for combining observations and multiple model output for an improved estimate of the global surface ozone distribution. *Geoscientific Model Development*, 12(3): 955-978 [DOI: 10.5194/gmd-12-955-2019]
- Charantonis A A, Badran F and Thiria S. 2015. Retrieving the evolution of vertical profiles of Chlorophyll-a from satellite observations using Hidden Markov Models and Self-Organizing Topological Maps. *Remote Sensing of Environment*, 163: 229-239 [DOI: 10.1016/j.rse.2015.03.019]
- Chen L F, Zhang Y, Zou M M, Xu Q, Li L J, Li X Y and Tao J H. 2015. Overview of atmospheric CO<sub>2</sub> remote sensing from space. *Journal of Remote Sensing*, 19(1): 1-11 (陈良富, 张莹, 邹铭敏, 徐谦, 李令军, 李小英, 陶金花. 2015. 大气 CO<sub>2</sub> 浓度卫星遥感进展. *遥感学报*, 19(1): 1-11) [DOI: 10.11834/jrs.20153331]
- Chen W, Zhao J, Cao C X and Tian H J. 2018. Shrub biomass estimation in semi-arid sandland ecosystem based on remote sensing technology. *Global Ecology and Conservation*, 16: e00479 [DOI: 10.1016/J.GECCO.2018.E00479]
- Chen X H, Feng K R, Liu N Y, Lu Y F, Tong Z Y, Ni B B, Liu Z A and Lin N. 2020. RainNet: a large-scale dataset for spatial precipitation downscaling. arXiv: 2012.09700
- Chen Y Z, Fu B J and Feng X M. 2017. Overview and outlook of remote sensing inversion of vegetation nitrogen content. *Acta Ecologica Sinica*, 37(18): 6240-6252 (陈永喆, 傅伯杰, 冯晓明. 2017. 遥感反演植被含氮量研究进展. *生态学报*, 37(18): 6240-6252) [DOI: 10.5846/stxb201707131271]
- Chen Z X, Ren J Q, Tang H J, Shi Y, Leng P, Liu J, Wang L M, Wu W B, Yao Y M and Hasiyuya. 2016. Progress and perspectives on agricultural remote sensing research and applications in China. *Journal of Remote Sensing*, 20(5): 748-767 (陈仲新, 任建强, 唐华俊, 史云, 冷佩, 刘佳, 王利民, 吴文斌, 姚艳敏, 哈斯图亚. 2016. 农业遥感研究应用进展与展望. *遥感学报*, 20(5): 748-767) [DOI: 10.11834/jrs.20166214]
- Cintra R S and de Campos Velho H F. 2014. Data assimilation by artificial neural networks for an atmospheric general circulation model: conventional observation. arXiv:1407.4360
- Combal B, Baret F, Weiss M, Trubuil A, Macé D, Pragnère A, Myneni R, Knyazikhin Y and Wang L. 2003. Retrieval of canopy biophysical variables from bidirectional reflectance: using prior information to solve the ill-posed inverse problem. *Remote Sensing of Environment*, 84(1): 1-15 [DOI: 10.1016/S0034-4257(02)00035-4]
- Davis D T, Chen Z, Tsang L, Hwang J N and Chang A T C. 1993. Retrieval of snow parameters by iterative inversion of a neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 31(4): 842-852 [DOI: 10.1109/36.239907]
- de Bezenac E, Pajot A and Gallinari P. 2018. Deep learning for physical processes: incorporating prior scientific knowledge. arXiv: 1711.07970
- De Jong K A. 1975. *Analysis of the Behavior of a Class of Genetic Adaptive Systems*. Ann Arbor, MI, United States: University of Michigan
- Dong L C, Li S W, Yang J, Shi W X and Zhang L. 2020. Investigating the performance of satellite-based models in estimating the surface PM<sub>2.5</sub> over China. *Chemosphere*, 256: 127051 [DOI: 10.1016/j.chemosphere.2020.127051]
- Dueben P D and Bauer P. 2018. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10): 3999-4009 [DOI: 10.

- 5194/gmd-11-3999-2018]
- Ebert-Uphoff I and Hilburn K. 2020. Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, 101(12): E2149-E2170 [DOI: 10.1175/BAMS-D-20-0097.1]
- Edwards J R and Parry M E. 1993. On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal*, 36(6): 1577-1613 [DOI: 10.5465/256822]
- El Hourany R, Abboud-Abi Saab M, Faour G, Aumont O, Crépon M and Thiria S. 2019. Estimation of secondary phytoplankton pigments from satellite observations using self-organizing maps (SOMs). *Journal of Geophysical Research: Oceans*, 124(2): 1357-1378 [DOI: 10.1029/2018JC014450]
- Elnaggar A A and Noller J S. 2009. Application of remote-sensing data and decision-tree analysis to mapping salt-affected soils over large areas. *Remote Sensing*, 2(1): 151-165 [DOI: 10.3390/rs2010151]
- Eyre J R, English S J and Forsythe M. 2020. Assimilation of satellite data in numerical weather prediction. Part I: the early years. *Quarterly Journal of the Royal Meteorological Society*, 146(726): 49-68 [DOI: 10.1002/qj.3654]
- Filippi A M, Güneralp İ and Randall J. 2014. Hyperspectral remote sensing of aboveground biomass on a river meander bend using multivariate adaptive regression splines and stochastic gradient boosting. *Remote Sensing Letters*, 5(5): 432-441 [DOI: 10.1080/215704X.2014.915070]
- Foody G M. 2008. RVM-based multi-class classification of remotely sensed data. *International Journal of Remote Sensing*, 29(6): 1817-1823 [DOI: 10.1080/01431160701822115]
- Friedman J H. 1991. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1): 1-67 [DOI: 10.1214/AOS/1176347963]
- Gao F, Wang Y P and Hu X Y. 2019. Evaluation of the suitability of Landsat, MERIS, and MODIS for identifying spatial distribution patterns of total suspended matter from a self-organizing map (SOM) perspective. *Catena*, 172: 699-710 [DOI: 10.1016/j.catena.2018.09.031]
- Gastellu-Etcheberry J P, Demarez V, Pinel V, and Zagolski F. 1996. Modeling radiative transfer in heterogeneous 3-D vegetation canopies. *Remote Sensing of Environment*, 58(2): 131-156 [DOI: 10.1016/0034-4257(95)00253-7]
- Geer A J. 2021. Learning earth system models from observations: machine learning or data assimilation?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194): 20200089 [DOI: 10.1098/rsta.2020.0089]
- Gentine P, Pritchard M, Rasp S, Reinaudi G and Yacalis G. 2018. Could machine learning break the convection parameterization deadlock?. *Geophysical Research Letters*, 45(11): 5742-5751 [DOI: 10.1029/2018GL078202]
- Ghulam A, Porton I and Freeman K. 2014. Detecting subcanopy invasive plant species in tropical rainforest by integrating optical and microwave (InSAR/PoInSAR) remote sensing data, and a decision tree algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 88: 174-192 [DOI: 10.1016/j.isprsjprs.2013.12.007]
- Goldberg D E and Holland J H. 1988. Genetic algorithms and machine learning. *Machine Learning*, 3: 95-99 [DOI: 10.1023/A:1022602019183]
- Gong P. 2009. Some essential questions in remote sensing science and technology. *Journal of Remote Sensing*, 13(1): 13-23 (宫鹏. 2009. 遥感科学与技术中的一些前沿问题. *遥感学报*, 13(1): 13-23)
- Grody N C and Basist A N. 1996. Global identification of snowcover using SSM/I measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 34(1): 237-249 [DOI: 10.1109/36.481908]
- Grossman Y L, Ustin S L, Jacquemoud S, Sanderson E W, Schmuck G and Verdebout J. 1996. Critique of stepwise multiple linear regression for the extraction of leaf biochemistry information from leaf reflectance data. *Remote Sensing of Environment*, 56(3): 182-193 [DOI: 10.1016/0034-4257(95)00235-9]
- Guo Y X, Tang Q H, Gong D Y and Zhang Z Y. 2017. Estimating ground-level PM<sub>2.5</sub> concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. *Remote Sensing of Environment*, 198: 140-149 [DOI: 10.1016/j.rse.2017.06.001]
- Hang R L, Liu Q S, Song H H, Sun Y B, Zhu F P and Pei H C. 2017. Graph regularized nonlinear ridge regression for remote sensing data analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(1): 277-285 [DOI: 10.1109/JSTARS.2016.2574802]
- Healey S P, Cohen W B, Yang Z Q, Brewer C K, Brooks E B, Gorelick N, Hernandez A J, Huang C Q, Hughes M J, Kennedy R E, Loveland T R, Moisen G G, Schroeder T A, Stehman S V, Vogelmann J E, Woodcock C E, Yang L M and Zhu Z. 2018. Mapping forest change using stacked generalization: an ensemble approach. *Remote Sensing of Environment*, 204: 717-728 [DOI: 10.1016/j.rse.2017.09.029]
- Hoerl A E and Kennard R W. 1970. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1): 69-82 [DOI: 10.1080/00401706.1970.10488635]
- Holland J H. 1975. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge: MIT Press
- Huang H, Ji X L, Xia F, Huang S H, Shang X, Chen H, Zhang M H, Dahlgren R A and Mei K. 2020. Multivariate adaptive regression splines for estimating riverine constituent concentrations. *Hydrological Processes*, 34(5): 1213-1227 [DOI: 10.1002/hyp.13669]
- Hyypää J, Hyypää H, Inkinen M, Engdahl M, Linko S and Zhu Y H. 2000. Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. *Forest Ecology and Management*, 128(1/2): 109-120 [DOI: 10.1016/S0378-1127(99)

- 00278-9]
- Irrgang C, Saynisch-Wagner J and Thomas M. 2020. Machine learning-based prediction of spatiotemporal uncertainties in global wind velocity reanalyses. *Journal of Advances in Modeling Earth Systems*, 12(5): e2019MS001876 [DOI: 10.1029/2019MS001876]
- Islam T, Srivastava P K, Dai Q, Gupta M and Zhuo L. 2015. Rain rate retrieval algorithm for conical-scanning microwave imagers aided by random forest, RReliefF, and multivariate adaptive regression splines (RAMARS). *IEEE Sensors Journal*, 15(4): 2186-2193 [DOI: 10.1109/JSEN.2014.2372814]
- Jacquemoud S and Baret F. 1990. PROSPECT: a model of leaf optical properties spectra. *Remote Sensing of Environment*, 34(2): 75-91 [DOI: 10.1016/0034-4257(90)90100-Z]
- Jia X W, Karpatne A, Willard J, Steinbach M, Read J, Hanson P C, Dugan H A and Kumar V. 2018. Physics guided recurrent neural networks for modeling dynamical systems: application to monitoring water temperature and quality in lakes. arXiv: 1810.02880
- Jiang C M, Esmailzadeh S, Azizzadenesheli K, Kashinath K, Mustafa M, Tchelepi H A, Marcus P, Prabhat M and Anandkumar A. 2020b. MeshfreeFlowNet: a physics-constrained deep continuous space-time super-resolution framework//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. Atlanta, GA, USA: IEEE: 1-15 [DOI: 10.1109/SC41405.2020.00013]
- Jiang Q O, Xu L D, Sun S Y, Wang M L and Xiao H J. 2021. Retrieval model for total nitrogen concentration based on UAV hyper spectral remote sensing data and machine learning algorithms - A case study in the Miyun Reservoir, China. *Ecological Indicators*, 124: 107356 [DOI: 10.1016/j.ecolind.2021.107356]
- Jiang S J, Zheng Y and Solomatine D. 2020a. Improving AI system awareness of geoscience knowledge: symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, 47(13): e2020GL088229 [DOI: 10.1029/2020GL088229]
- Kang J, Jin R, Li X, Zhang Y and Zhu Z L. 2018. Spatial upscaling of sparse soil moisture observations based on ridge regression. *Remote Sensing*, 10(2): 192 [DOI: 10.3390/rs10020192]
- Keller C A and Evans M J. 2019. Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. *Geoscientific Model Development*, 12(3): 1209-1225 [DOI: 10.5194/gmd-12-1209-2019]
- Koelmeijer R B A, Homan C D and Matthijssen J. 2006. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmospheric Environment*, 40(27): 5304-5315 [DOI: 10.1016/j.atmosenv.2006.04.044]
- Kraft B, Jung M, Körner M and Reichstein M. 2020. Hybrid modeling: fusion of a deep learning approach and a physics-based model for global hydrological modeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020: 1537-1544 [DOI: 10.5194/isprs-archives-XLIII-B2-2020-1537-2020]
- Kukreja S L, Löfberg J and Brenner M J. 2006. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proceedings Volumes*, 39(1): 814-819 [DOI: 10.3182/20060329-3-AU-2901.00128]
- Kuter S, Akyurek Z and Weber G W. 2018. Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines. *Remote Sensing of Environment*, 205: 236-252 [DOI: 10.1016/j.rse.2017.11.021]
- Lee H J, Liu Y, Coull B A, Schwartz J and Koutrakis P. 2011. A novel calibration approach of MODIS AOD data to predict PM<sub>2.5</sub> concentrations. *Atmospheric Chemistry and Physics*, 11(15): 7991-8002 [DOI: 10.5194/acp-11-7991-2011]
- Lee Y, Han D, Ahn M H, Im J and Lee S J. 2019. Retrieval of total precipitable water from Himawari-8 AHI data: a comparison of random forest, extreme gradient boosting, and deep neural network. *Remote Sensing*, 11(15): 1741 [DOI: 10.3390/rs11151741]
- Li D R, Yao Y and Shao Z F. 2014a. Big data in smart city. *Geomatics and Information Science of Wuhan University*, 39(6): 631-640 (李德仁, 姚远, 邵振峰. 2014a. 智慧城市中的大数据. 武汉大学学报(信息科学版), 39(6): 631-640) [DOI: 10.13203/j.whugis20140135]
- Li D R, Zhang L P and Xia G S. 2014b. Automatic analysis and mining of remote sensing big data. *Acta Geodaetica et Cartographica Sinica*, 43(12): 1211-1216 (李德仁, 张良培, 夏桂松. 2014b. 遥感大数据自动分析与数据挖掘. 测绘学报, 43(12): 1211-1216) [DOI: 10.13485/jc.nki.1-20892.0140.187]
- Li L, Cheng Y B, Ustin S, Hu X T and Riaño D. 2008. Retrieval of vegetation equivalent water thickness from reflectance using genetic algorithm (GA)-partial least squares (PLS) regression. *Advances in Space Research*, 41(11): 1755-1763 [DOI: 10.1016/j.asr.2008.02.015]
- Li L, Ustin S L and Riano D. 2007. Retrieval of fresh leaf fuel moisture content using genetic algorithm partial least squares (GA-PLS) modeling. *IEEE Geoscience and Remote Sensing Letters*, 4(2): 216-220 [DOI: 10.1109/LGRS.2006.888847]
- Li T W, Shen H F, Yuan Q Q and Zhang L P. 2020a. Geographically and temporally weighted neural networks for satellite-based mapping of ground-level PM<sub>2.5</sub>. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167: 178-188 [DOI: 10.1016/j.isprsjrs.2020.06.019]
- Li T W, Shen H F, Zeng C, Yuan Q Q and Zhang L P. 2017. Point-surface fusion of station measurements and satellite observations for mapping PM<sub>2.5</sub> distribution in China: methods and assessment. *Atmospheric Environment*, 152: 477-489 [DOI: 10.1016/j.atmosenv.2017.01.004]
- Li X, Liu F and Fang M. 2020b. Harmonizing models and observations: data assimilation in earth system science. *Science China Earth Sciences*, 63(8): 1059-1068 [DOI: 10.1007/s11430-019-9620-x]

- Li X W. 2005. Retrospect, prospect and innovation in quantitative remote sensing. *Journal of Henan University (Natural Science)*, 35(4): 49-56 (李刘文. 2005. 定量遥感的发展与创新. *河南大学学报(自然科学版)*, 35(4): 49-56) [DOI: 10.15991/j.cnki.411100.2005.04.012]
- Liang S L, Cheng J, Jia K, Jiang B, Liu Q, Liu S H, Xiao Z Q, Xie X H, Yao Y J, Yuan W P, Zhang X T and Zhao X. 2016. Recent progress in land surface quantitative remote sensing. *Journal of Remote Sensing*, 20(5): 875-898 (梁顺林, 程洁, 贾坤, 江波, 刘强, 刘素红, 肖志强, 谢先红, 姚云军, 袁文平, 张晓通, 赵祥. 2016. 陆表定量遥感反演方法的发展新动态. *遥感学报*, 20(5): 875-898) [DOI: 10.11834/jrs.20166258]
- Liang S L, Li X W and Wang J D. 2013. *Quantitative Remote Sensing: Concepts and Algorithms*. Beijing: Science Press (梁顺林, 李刘文, 王锦地. 2013. 定量遥感: 理念与算法. 北京: 科学出版社)
- Liu L, Weng C S, Li S L, Husi L, Hu S and Dong P Y. 2021. Passive remote sensing of ice cloud properties at terahertz wavelengths based on genetic algorithm. *Remote Sensing*, 13(4): 735 [DOI: 10.3390/rs13040735]
- Lyu B, Hu Y T, Zhang W X, Du Y S, Luo B, Sun X L, Sun Z, Deng Z, Wang X J, Liu J, Wang X S and Russell A G. 2019. Fusion method combining ground-level observations with chemical transport model predictions using an ensemble deep learning framework: application in China to estimate spatiotemporally-resolved PM<sub>2.5</sub> exposure fields in 2014-2017. *Environmental Science and Technology*, 53(13): 7306-7315 [DOI: 10.1021/acs.est.9b01117]
- Ma J, Ding Y X, Cheng J C P, Jiang F F and Wan Z W. 2019. A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM<sub>2.5</sub>. *Journal of Cleaner Production*, 237: 117729 [DOI: 10.1016/j.jclepro.2019.117729]
- Ma Y C, Zhang Z, Kang Y H and Özdoğan M. 2021. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, 259: 112408 [DOI: 10.1016/j.rse.2021.112408]
- Malmgren-Hansen D, Laparra V, Aasbjerg Nielsen A and Camps-Valls G. 2019. Statistical retrieval of atmospheric profiles with deep convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158: 231-240 [DOI: 10.1016/j.isprsjrs.2019.10.002]
- Mo S X, Zhu Y H, Zabarans N, Shi X Q and Wu J C. 2019. Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. *Water Resources Research*, 55(1): 703-728 [DOI: 10.1029/2018wr023528]
- Monteith J L. 1965. Evaporation and environment. *Symposium of the Society of Exploratory Biology*, 19: 205-234
- Mountrakis G, Im J and Ogole C. 2011. Support vector machines in remote sensing: a review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3): 247-259 [DOI: 10.1016/j.isprsjrs.2010.11.001]
- Pal A, Mahajan S and Norman M R. 2019. Using deep neural networks as cost-effective surrogate models for super-parameterized E3SM radiative transfer. *Geophysical Research Letters*, 46(11): 6069-6079 [DOI: 10.1029/2018GL081646]
- Parikh J A, DaPonte J S, Damodaran M and Sherman P. 1990. Application of neural networks to pattern recognition problems in remote sensing and medical imagery//*Proceedings Volume 1294, Applications of Artificial Neural Networks*. Orlando, FL, United States: SPIE [DOI: 10.1117/12.21165]
- Pathak J, Wikner A, Fussell R, Chandra S, Hunt B R, Girvan M and Ott E. 2018. Hybrid forecasting of chaotic processes: using machine learning in conjunction with a knowledge-based model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(4): 041101 [DOI: 10.1063/1.5028373]
- Pham T D, Yokoya N, Xia J S, Ha N T, Le N N, Nguyen T T T, Dao T H, Vu T T P, Pham T D and Takeuchi W. 2020. Comparison of machine learning methods for estimating mangrove above-ground biomass using multiple source remote sensing data in the red river delta biosphere reserve, Vietnam. *Remote Sensing*, 12(8): 1334 [DOI: 10.3390/rs12081334]
- Priestley C H B and Taylor R J. 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, 100(2): 81-92 [DOI: 10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2]
- Qin K, Rao L L, Xu J, Bai Y, Zou J H, Hao N, Li S S and Yu C. 2017. Estimating ground level NO<sub>2</sub> concentrations over central-eastern China using a satellite-based geographically and temporally weighted regression model. *Remote Sensing*, 9(9): 950 [DOI: 10.3390/rs9090950]
- Qu Y H, Wang J D, Wan H W, Li X W and Zhou G Q. 2008. A Bayesian network algorithm for retrieving the characterization of land surface vegetation. *Remote Sensing of Environment*, 112(3): 613-622 [DOI: 10.1016/j.rse.2007.03.031]
- Rasp S, Pritchard M S and Gentile P. 2018. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, 115(39): 9684-9689 [DOI: 10.1073/pnas.1810286115]
- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N and Prabhat N. 2019. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743): 195-204 [DOI: 10.1038/s41586-019-0912-1]
- Ruckstuhl Y, Janjić T and Rasp S. 2021. Training a convolutional neural network to conserve mass in data assimilation. *Nonlinear Processes in Geophysics*, 28(1): 111-119 [DOI: 10.5194/npg-28-111-2021]
- Sawada Y. 2020. Machine learning accelerates parameter optimization and uncertainty assessment of a land surface model. *Journal of Geophysical Research: Atmospheres*, 125(20): e2020JD032688

- [DOI: 10.1029/2020JD032688]
- Sawaya K E, Olmanson L G, Heinert N J, Brezonik P L and Bauer M E. 2003. Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery. *Remote Sensing of Environment*, 88(1/2): 144-156 [DOI: 10.1016/j.rse.2003.04.006]
- Shen H F, Li T W, Yuan Q Q and Zhang L P. 2018. Estimating regional ground-level  $PM_{2.5}$  directly from satellite top-of-atmosphere reflectance using deep belief networks. *Journal of Geophysical Research: Atmospheres*, 123(24): 13875-13886 [DOI: 10.1029/2018JD028759]
- Sonderby C K, Espeholt L, Heck J, Dehghani M, Oliver A, Salimans T, Agrawal S, Hickey J and Kalchbrenner N. 2020. MetNet: a neural weather model for precipitation forecasting. arXiv: 2003.12140
- Su T N, Laszlo I, Li Z Q, Wei J and Kalluri S. 2020. Refining aerosol optical depth retrievals over land by constructing the relationship of spectral surface reflectances through deep learning: application to Himawari-8. *Remote Sensing of Environment*, 251: 112093 [DOI: 10.1016/j.rse.2020.112093]
- Tai A P K, Mickley L J and Jacob D J. 2010. Correlations between fine particulate matter ( $PM_{2.5}$ ) and meteorological variables in the United States: implications for the sensitivity of  $PM_{2.5}$  to climate change. *Atmospheric Environment*, 44(32): 3976-3984 [DOI: 10.1016/j.atmosenv.2010.06.060]
- Tedesco M, Pulliainen J, Takala M, Hallikainen M and Pampaloni P. 2004. Artificial neural network-based techniques for the retrieval of SWE and snow depth from SSM/I data. *Remote Sensing of Environment*, 90(1): 76-85 [DOI: 10.1016/j.rse.2003.12.002]
- Tseng M H, Chen S J, Hwang G H and Shen M Y. 2008. A genetic algorithm rule-based approach for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(2): 202-212 [DOI: 10.1016/j.isprsjprs.2007.09.001]
- Tuia D, Roscher R, Wegner J D, Jacobs N, Zhu X X and Camps-Valls G. 2021. Toward a collective agenda on AI for earth science data analysis. *IEEE Geoscience and Remote Sensing Magazine*, 9(2): 88-104 [DOI: 10.1109/MGRS.2020.3043504]
- Uyanik G K and Güler N. 2013. A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, 106, 234-240.
- Van der Tol C, Verhoef W, Timmermans J, Verhoef A and Su Z. 2009. An integrated model of soil-canopy spectral radiances, photosynthesis, fluorescence, temperature and energy balance. *Biogeosciences*, 6(12): 3109-3129 [DOI: 10.5194/bg-6-3109-2009]
- van Donkelaar A, Martin R V, Spurr R J D and Burnett R T. 2015. High-resolution satellite-derived  $PM_{2.5}$  from optimal estimation and geographically weighted regression over North America. *Environmental Science and Technology*, 49(17): 10482-10491 [DOI: 10.1021/acs.est.5b02076]
- Vapnik V, Golowich S E and Smola A. 1996. Support vector method for function approximation, regression estimation and signal processing//Proceedings of the 9th International Conference on Neural Information Processing Systems. Denver, Colorado: MIT Press: 281-287
- Veerman M A, Pincus R, Stoffer R, van Leeuwen C M, Podareanu D and van Heerwaarden C C. 2021. Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194): 20200095 [DOI: 10.1098/rsta.2020.0095]
- Verrelst J, Camps-Valls G, Muñoz-Mari J, Rivera J P, Veroustraete F, Clevers J G P W and Moreno J. 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties - A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108: 273-290 [DOI: 10.1016/j.isprsjprs.2015.05.005]
- von Rueden L, Mayer S, Sifa R, Bauckhage C and Garecke J. 2020. Combining machine learning and simulation to a hybrid modeling approach: current and future directions//18th International Symposium on Intelligent Data Analysis. Konstanz, Germany: Springer: 548-560 [DOI: 10.1007/978-3-030-44584-3\_43]
- Wang H Q, Magagi R, Goita K, Trudel M, McNairn H and Powers J. 2019. Crop phenology retrieval via polarimetric SAR decomposition and Random Forest algorithm. *Remote Sensing of Environment*, 231: 111234 [DOI: 10.1016/j.rse.2019.111234]
- Wang J and Christopher S A. 2003. Intercomparison between satellite-derived aerosol optical thickness and  $PM_{2.5}$  mass: implications for air quality studies. *Geophysical Research Letters*, 30(21): 2095 [DOI: 10.1029/2003GL018174]
- Wang L L, Hunt E R, Qu J J, Hao X J and Daughtry C S T. 2013. Remote sensing of fuel moisture content from ratios of narrow-band vegetation water and dry-matter indices. *Remote Sensing of Environment*, 129: 103-110 [DOI: 10.1016/j.rse.2012.10.027]
- Wei C W, Huang J F, Mansaray L R, Li Z H, Liu W W and Han J H. 2017. Estimation and mapping of winter oilseed rape LAI from high spatial resolution satellite data based on a hybrid method. *Remote Sensing*, 9(5): 488 [DOI: 10.3390/rs9050488]
- Wu J L, Kashinath K, Albert A, Chirila D, Prabhat N and Xiao H. 2020. Enforcing Statistical Constraints in Generative Adversarial Networks for Modeling Chaotic Dynamical Systems. *Journal of Computational Physics*, 406: 109209 [DOI: 10.1016/j.jcp.2019.109209]
- Wu S S, Wang Z Y, Du Z H, Huang B, Zhang F and Liu R Y. 2021. Geographically and temporally neural network weighted regression for modeling spatiotemporal non-stationary relationships. *International Journal of Geographical Information Science*, 35(3): 582-608 [DOI: 10.1080/13658816.2020.1775836]
- Xu G H. 1997. Progress and prospects of information science of remote sensing. *World Sci-Tech R and D*, 19(3): 32-40 (徐冠华. 1997. 遥感信息科学的进展和展望. 世界科技研究与发展, 19(3): 32-40)



- Xu X Q, Lu J S, Zhang N, Yang T C, He J Y, Yao X, Cheng T, Zhu Y, Cao W X and Tian Y C. 2019. Inversion of rice canopy chlorophyll content and leaf area index based on coupling of radiative transfer and Bayesian network models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150: 185-196 [DOI: 10.1016/j.isprsjprs.2019.02.013]
- Yang S Y, Yang D W, Chen J S, Santisirisomboon J, Lu W W and Zhao B X. 2020. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. *Journal of Hydrology*, 590: 125206 [DOI: 10.1016/j.jhydrol.2020.125206]
- Yin J N, Medellin-Azuara J, Escrivá-Bou A and Liu Z. 2021. Bayesian machine learning ensemble approach to quantify model uncertainty in predicting groundwater storage change. *Science of the Total Environment*, 769: 144715 [DOI: 10.1016/j.scitotenv.2020.144715]
- Yu F, Zhao Y S and Li H T. 2012. Soil moisture retrieval based on GA-BP neural networks algorithm. *Journal of Infrared and Millimeter Waves*, 31(3): 283-288 (余凡, 赵英时, 李海涛. 2012. 基于遗传BP神经网络的主被动遥感协同反演土壤水分. *红外与毫米波学报*, 31(3): 283-288) [DOI: 10.3724/SP.J.1010.2012.00283]
- Yuan Q Q, Shen H F, Li T W, Li Z W, Li S W, Jiang Y, Xu H Z, Tan W W, Yang Q Q, Wang J W, Gao J H and Zhang L P. 2020. Deep learning in environmental remote sensing: achievements and challenges. *Remote Sensing of Environment*, 241: 111716 [DOI: 10.1016/J.RSE.2020.111716]
- Yuval J and O'Gorman P A. 2020. Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, 11(1): 3295 [DOI: 10.1038/s41467-020-17142-3]
- Zanna L, Brankart J M, Huber M, Leroux S, Penduff T and Williams P D. 2019. Uncertainty and scale interactions in ocean ensembles: from seasonal forecasts to multidecadal climate predictions. *Quarterly Journal of the Royal Meteorological Society*, 145(S1): 160-175 [DOI: 10.1002/qj.3397]
- Zhang B. 2018. Remotely sensed big data era and intelligent information extraction. *Geomatics and Information Science of Wuhan University*, 43(12): 1861-1871 (张兵. 2018. 遥感大数据时代与智能信息提取. *武汉大学学报(信息科学版)*, 43(12): 1861-1871) [DOI: 10.13203/j.whugis20180172]
- Zhang Y and Li Z Q. 2015. Remote sensing of atmospheric fine particulate matter (PM<sub>2.5</sub>) mass concentration near the ground from satellite observation. *Remote Sensing of Environment*, 160: 252-262 [DOI: 10.1016/j.rse.2015.02.005]
- Zhang Y Z, Qu Y H, Wang J D, Liang S L and Liu Y. 2012. Estimating leaf area index from MODIS and surface meteorological data using a dynamic Bayesian network. *Remote Sensing of Environment*, 127: 30-43 [DOI: 10.1016/j.rse.2012.08.015]
- Zhao W L, Gentile P, Reichstein M, Zhang Y, Zhou S, Wen Y Q, Lin C J, Li X and Qiu G Y. 2019a. Physics-constrained machine learning of evapotranspiration. *Geophysical Research Letters*, 46(24): 14496-14507 [DOI: 10.1029/2019GL085291]
- Zhu J Y, Tan S R, King J, Derksen C, Lemmetyinen J and Tsang L. 2018. Forward and inverse radar modeling of terrestrial snow using SnowSAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 56(12): 7122-7132 [DOI: 10.1109/TGRS.2018.2848642]
- Zou H and Hastie T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301-320 [DOI: 10.1111/j.1467-9868.2005.00503.x]

## Research progress and challenges of data-driven quantitative remote sensing

YANG Qianqian<sup>1</sup>, JIN Caiyi<sup>1</sup>, LI Tongwen<sup>2</sup>, YUAN Qiangqiang<sup>1</sup>, SHEN Huanfeng<sup>3</sup>, ZHANG Liangpei<sup>4</sup>

1. School of Geodesy and Geomatics, Wuhan University, Wuhan 430079;

2. School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai 519082;

3. School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079;

4. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079

**Abstract:** Quantitative remote sensing is a technique for quantitatively inferring or inverting earth environmental variable from the original remote sensing observations, it is an important step to turn electromagnetic wave signals into easy-to-understand information about surface environment. Traditional quantitative remote sensing methods are mainly model-driven, emphasizing inversion through mathematical or physical models. With the development and popularization of artificial intelligence technology, data-driven methods have gradually received widespread attention. It emphasizes the use of machine learning methods to mine the information contained in remote sensing observation data to achieve the quantitative inversion of geophysical parameters. With the support of powerful computing capacity, the data-driven method has achieved gratifying achievements in many fields of quantitative remote sensing. This article systematically summarizes the

principles, characteristics, and applications in quantitative remote sensing field of different types of data-driven models, including regression algorithms, regularization methods, instance-based algorithms, decision tree, Bayesian methods, kernel based algorithms, genetic algorithms, ensemble learning, artificial neural network, and deep learning. Though data-driven models show satisfying retrieval performance in multiple fields, its drawbacks of ignoring the laws of physics and lack of causality have also brought resistance to its development. In this context, coupling the laws of physics and machine learning to develop an inversion framework driven by both models and data has become a new research hotspot. Some pioneering researches have already achieved delightful performance through using machine learning to assist physical models or restricting machine learning with physical laws. Using machine learning techniques to optimize the systematic basis, the sub-model, and the model parameters largely improve the performance of model-driven methods. Meanwhile, integrating physics knowledge into machine learning models through adjusting the training data, modifying the loss function, and constraining the solution space also benefit the improvement of data-driven models. However, there are still great challenges to be broken through. Physical models contain complex mechanisms and rich knowledge, current fusion of data-driven and model-driven methods are quite shallow with very limited amount of physical knowledge being used. A deeper coupling strategy is worth exploring in the future. Besides, the uncertainty, generalization, and transferability of the joint model have not been scientifically evaluated currently, to which attention should be paid. Finally, there are many cases when the training samples were very difficult to obtain, therefore, the applicability of the joint model in the case of small samples is also a problem that needs to be solved urgently. The deep, robust, and generalizable coupling of data-driven and model-driven models is expected in the future.

**Key words:** remote sensing, quantitative remote sensing, model-driven, data-driven, deep learning, machine learning, fusion

**Supported by** National Natural Science Foundation of China (No. 41922008); Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA19090104)