光伏预测中的机器学习算法简介

摘要：

预测太阳系统的输出功率是为了电网的良好运行，或者是为了对发生在太阳系中的能量流进行优化管理。在预测太阳系统的产出之前，必须把预测的重点放在太阳辐射上。全球太阳辐射预测可以通过几种方法进行;这两大类是利用云图像结合物理模型，以及机器学习模型。在此背景下，本文的目的是对利用机器学习方法进行太阳辐照的预测方法进行概述。虽然，许多论文描述了神经网络或支持向量回归等方法，但它将显示其他方法(回归树、随机森林、梯度增强等)在预测的背景下开始使用。由于数据集的多样性、时间步长、预测范围、设置和性能指标的多样性，使得该方法的性能排序更加复杂。总的来说，预测误差是相对的。为了提高预测性能，一些作者提出了使用混合模型或采用集成预测方法。

1、简介：

电气操作人员应在任何时刻确保电力生产与消耗之间的精确平衡。但这通常很难在传统的、可控的能源生产系统维持，主要是在小的或不相连的(孤立的)电网(如在岛屿上发现)。现在许多国家都在考虑使用可再生能源发电，这就造成了更多的问题，因为资源(太阳辐射，风等)本身存在不稳定性。因此，能够有效地预测太阳辐射显得非常重要，特别是在高能集成的情况下。

1.1光伏光能生产的预测至关重要

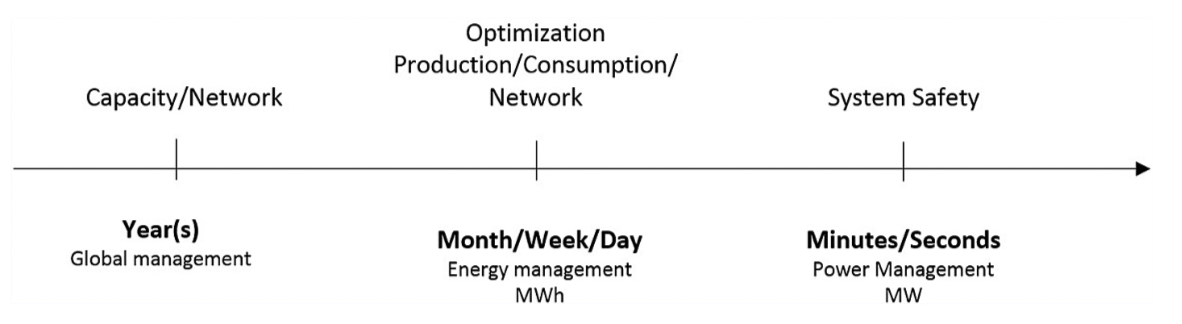
对近期全球能源供应来说，最重要的挑战之一是将可再生能源(特别是不可预测的风能和太阳能)大规模集成到现有或未来的能源供应结构中。电气操作人员应在任何时刻确保电力生产与消耗之间的精确平衡。事实上，操作者在维持这种平衡的过程中往往会遇到一些困难，这些困难主要集中在小型或不相连的电网(如在岛屿上)。电力系统的可靠性取决于系统能否适应预期的和意外的变化(生产和消费)以及干扰，同时保持服务质量和服务的连续性。然后，能源供应商必须在不同的时间范围内完成不同的管理任务（见图1）。

图1 电网中能量管理的预测尺度[2]

可再生能源与电网的集成，加剧了电网管理的复杂性和生产、消费平衡的复杂性，这是由于它们的间歇性和不可预测性[1,2]。太阳能生产的间歇性和非可控特性带来了许多其他问题，如电压波动、局部电能质量和稳定性问题[3,4]。因此，预测太阳系统的输出功率是为了电网的有效运行，或者是对发生在太阳系的能量流的最优管理[5]。也有必要对储量进行估算，对电力系统进行调度，进行拥塞管理，对随机生产的存储进行最优管理，并在电力市场上交易产生的电力，最终实现电力生产成本的降低[1,3,6,7]。由于太阳能发电的大量增加，太阳能产量的预测变得越来越重要[8]。为了避免在可再生电力生产中出现较大的变化，必须将系统运行的完整预测与存储解决方案相结合。各种存储系统正在开发中，它们是一种可行的解决方案，用于吸收这种系统产生的过剩的能量和能量(并在消耗高峰期释放它)，用于带来非常短的波动和保持电能质量的连续性。这些存储选项通常分为三类:

（1）大量的储能或能量管理存储介质被用来分离产生和消耗的时间；

（2）分布式发电或桥接电源——这种方法被用于高峰修剪，为了确保在能源修改期间服务的连续性，储存被使用几分钟到几个小时；

（3）电能质量存储的时间尺度约为几秒，仅用于保证终端使用电能质量的连续性。

表1展示了这三个类别及其技术规范。如图所示，在不同的情况下，每一种类型的存储都被用于解决不同的问题，都有着不同的时间范围和大量的能量。

表1显示，在许多情况下，电能存储可以被广泛用于许多场合以及一些应用，例如使用时间的函数和最终用户的电力需求。最后，它显示了如果需要在不同时间水平下的储能行为以及进行适当的管理就需要了解太阳系统在不同的范围内所产生的能量或者能源的知识:对于大容量能量储存,，比如说超短期或者短期每小时或每天的电能质量类别。同样,电气操作手需要知道未来在不同的时间范围例如从1到3天生产(图1),准备生产系统以及开始启动的发电厂(表2)的时间或长或短。开始一个发电厂，从开启液压的5分钟到核能的40h都有。此外，发电厂的电力增长有时很低，因此，为了在生产和消费之间取得有效的平衡，电力或新的生产的开始需要提前预期。

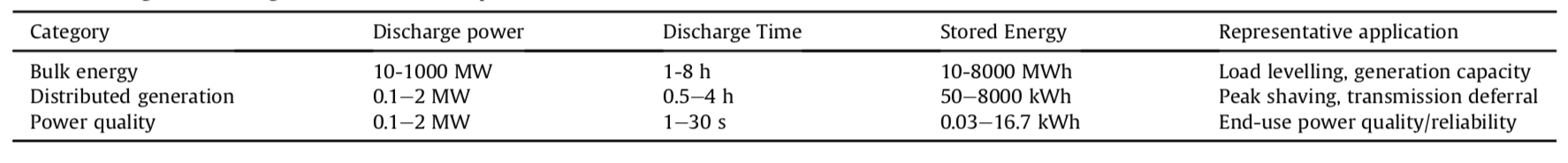
此外，预测期可以而且必须从5分钟到几天，由Diagne等[6]所证实的。Elliston和MacGill[10]概述了各种太阳能系统(PV、热能、集中式太阳能热电厂等)的太阳辐射预测的原因。因此，似乎很明显，预测数据的时间步可能因目标和预测范围而异。所有这些原因都表明了预测的重要性，无论是在生产中还是在能源消耗方面。对预测的需求导致了使用有效预测模型的必要性。下一节将介绍各种可用的预测方法。

表1 三类存储及其技术规格

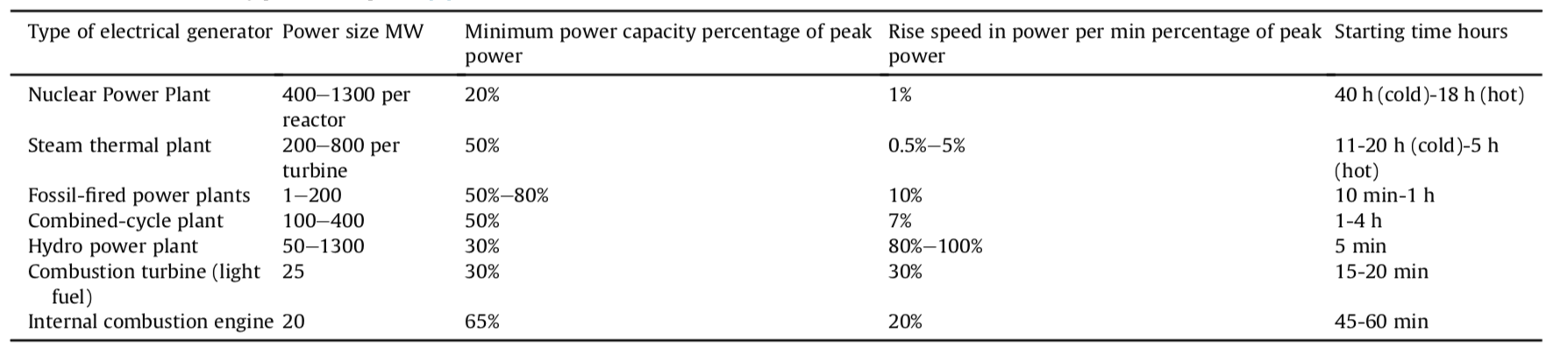


表2 电力生产厂的特点[9]

1.2 现有的预测方法

太阳能预测可以用几种方法进行;这两大类是云图像结合物理模型，以及机器学习模型。该方法的选择主要取决于预测水平;实际上，所有的模型在相同预测期上都没有相同的精确度表现。根据目标预测时间的不同，预测太阳辐照度的方法多种多样。文献将这些方法分为两类:

（1）外推法和统计过程。使用卫星图像或测量的地面和天空图之类通常适合6 h的短期预测。这个类可以分为两个子类,在很短的时间域内称为“临时预报”(0 -3 h),实时的预测是基于推断测量[5];在短期预测(3-6h)中，数值天气预报(NWP)模型与实时测量或卫星数据相结合的后处理模块[5,11]。

（2）NWP模型能够提前两天或超过[12,13]（提前6天[13]）预测。 这些NWP模型有时与后处理模块结合在一起，并且经常使用卫星信息[2]。

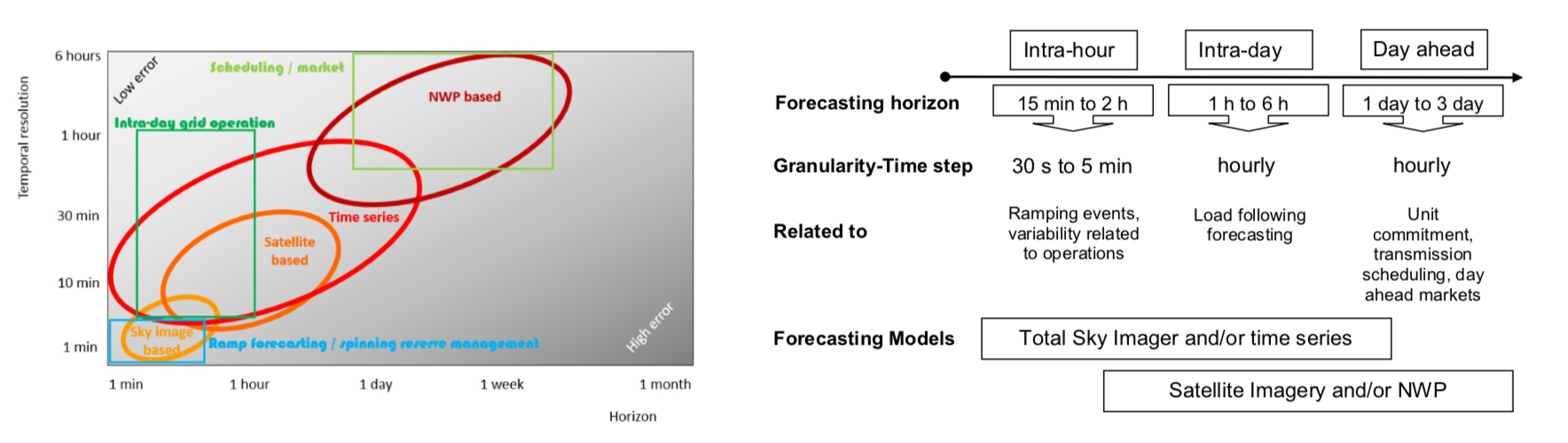
 图2a和b [6,14]总结了现有方法与预测时限，目标和时间步长。

图2 a）预测误差与预测模型（左）b）预测范围，预测模型和相关活动之间的关系（右）[6,14]

NWP模型预测局部云形成的概率，然后使用动态大气模型间接预测透射辐射。通过估算云的运动外推或统计模型并预测它们对未来的影响，分析全球辐射的历史时间序列，从卫星遥感[15]或地面测量[16]中分析[6,13,17]。混合方法可以改进所有这些方法的某些方面[6,14]。统计方法允许预测每小时太阳辐射（或在较低的时间步长），NWP模型使用解释变量（主要是云的运动和来自大气的方向）提前N步来预测全球辐射[15]。有关预测方法的很好的概述，以及它们的局限性和准确性，可以在参考文献 [1,5,6,10,12,14,18]中找到。进行基准研究以评估辐照度预测的准确性，并比较不同的预测方法[8,13,17,19-21]。而且，精度评估参数往往不同;通常使用一些参数，如相关系数和均方根误差，但并不总是适用于比较模型性能。因此用于评估准确度的时间段差别很大。其中一些人分析了一年或几年期间的模型精确度，而另一些人则在几周内引入了潜在的海洋偏差。在这些情况下，进行比较并不容易，如本文所示，所选结果的准确性必须在选择正确的使用方法时仔细评估。作为COST ACTION ES1002的一部分（欧洲科学和技术合作）[22]关于可再生能源天气情报（WIRE）的文献回顾了对适用于可再生能源系统（主要是太阳能和风能）的预测精度进行了研究。本文介绍了基于机器学习的可用于太阳辐射预测的各种方法。许多评论文章都是可用的，但很难找到完全致力于机器学习方法的论文，并且最近的一些预测模型如随机森林，增强树或回归树被整合在一起。在下一节中，介绍了文献中用于预测全局辐射的不同方法和用于估计模型性能的参数。

2、机器学习方法

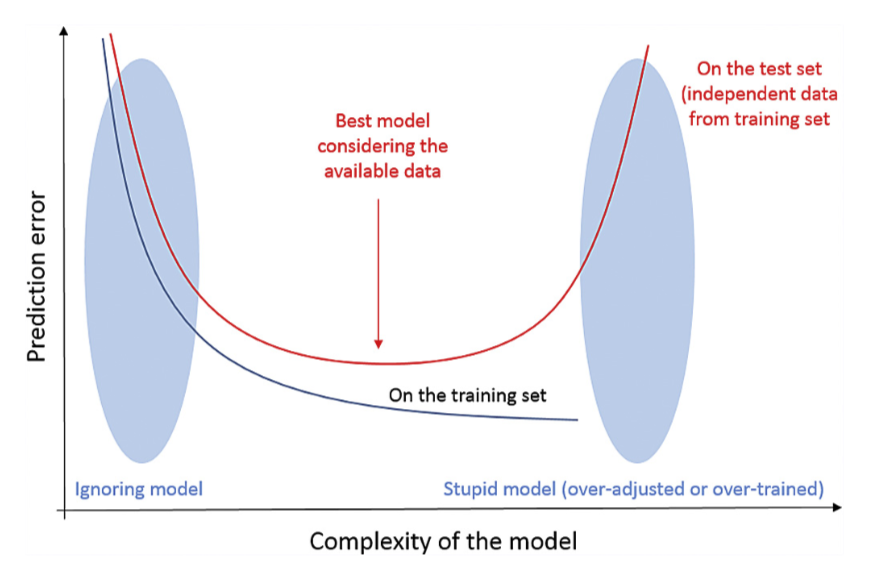
机器学习是计算机科学的一个子领域，它被分类为人工智能方法。它可以用于多个领域，这种方法的优点是一个模型可以解决不可能由显式算法表示的问题。在参考文献[23]读者可以找到一些关于太阳能预测的机器学习和确定性方法的详细回顾。即使表示是不可能的，机器学习模型也会找到输入和输出之间的关系;这种特性允许在许多情况下使用机器学习模型，例如模式识别，分类问题，垃圾邮件过滤以及数据挖掘和预测问题。分类和数据挖掘在这个领域特别有趣，因为人们必须使用大数据集，预处理和数据准备的任务可以通过机器学习模型进行。在此步骤之后，机器学习模型可用于预测问题。在全球水平辐照度预测中，模型可以以三种不同的方式使用[24]：

（1）基于其他气象和地理参数的结构模型;

（2）时间序列模型只考虑太阳辐射的历史观测数据作为输入特征（内生预测）;

（3）考虑太阳辐射和其他变量作为外生变量（外生预测）的混合模型。

如前所述，机器学习是人工智能的一个分支。 它涉及到可以从数据集中学习的系统的构建和研究，使计算机能够在没有明确编程的情况下学习。 在预测性学习问题中，系统由一个随机“输出”或“响应”变量y和一组随机“输入”或“解释性变量”组成。 使用“训练”样本， 已知（y，x）值，目标是获得函数f （x），书x到y的映射，使得关于联合分布函数的某些指定损失函数最小化，f（x）的期望值最小化 - 在所有的（y，x）上：

经常使用的损失函数; f（x）包括平方误差和绝对误差用于回归和负二项对数似然性的分类。 一个普通的过程是将f（x）限制为一个参数化函数类的成员，其中是一组有限的参数，其联合值 确定种类成员。 通常所有用于机器学习的方法，特别是监督案例，都面临着偏差 - 方差的折衷（见图3）。 这是试图最小化两个误差来源的问题，它们阻止了监督学习算法在其训练集之外进行泛化：

- 偏差是学习算法中出现的错误假设的偏差（错误）。 较高的偏见值可能会导致偏差算法失去了建立实际关系的能力和目标输出（安装不足）。

- 方差是实际捕获小的错误训练集中的波动。 应该注意，高方差会导致过拟合，导致训练数据集中的随机噪声建模，而不是预期的输出。偏差 - 方差关系的分解是一种调查学习算法对于特定问题的预期泛化误差的方法，该问题是三个项的总和，偏差，方差和不可约的误差，这是由问题中的噪音引起的本身。

在这一部分，我们介绍了用于预测的不同机器学习模型，最初是用于分类和数据准备的模型，其次是监督学习模型，第三是无监督学习模型，最后是集成学习模型。

图3 偏差方差权衡

2.1 分类和数据准备

机器学习算法从数据中学习。因此选择正确的数据并妥善准备以使问题得到有效解决至关重要。

2.1.1.主成分分析（PCA）

主成分分析（PCA）是一种统计方法，它使用正交变换将可能相关变量的一组观测值转换成一组称为主成分的线性不相关变量的值[25]。在此过程中创建的主要组件的数量低于或等于原始变量的数量。这样的变换以这样的方式定义：第一主成分具有可能的最大变异，即为了说明数据中的最大变异性，并且每个后续成分在其正交于之前的组件。结果，所得到的矢量形成不相关的正交基集。应该注意的是，主成分是正交的，因为它们是对称的协方差矩阵的特征向量。此外，PCA对原始变量的相对标度敏感[26]。

2.1.2.朴素贝叶斯分类和贝叶斯网络

在机器学习中，朴素贝叶斯分类器是一组简单的概率分类器，它基于贝叶斯定理和特征之间强大（天真）的独立性假设。朴素贝叶斯分类器具有高度可扩展性，需要许多参数与学习问题中的变量数量（特征/预测变量）成比例。最大似然训练可以通过评估一个需要线性时间的闭形式表达式来完成，而不是像许多其他类型的分类器那样使用昂贵的迭代近似法[27]。贝叶斯网络也称为贝叶斯网络，贝叶斯模型，信念网络或概率有向无环图模型是概率图模型，其是一种统计模型，其表示一组随机变量及其条件依赖关系通过有向无环图（DAG）。

2.1.3。数据挖掘方法

数据挖掘包括在大数据集中发现有趣的，意外的或有价值的结构，这些结构可以被称为slogan Big Data[28]。换句话说，数据挖掘包括从一个非常大的数据集中提取最重要的信息。确实，经典的统计推断已经开发出来用于处理小样本。在存在非常大的数据库的情况下，所有标准统计指标都变得重要且有趣（例如，对于100万数据，相关系数的显着性阈值非常低，达到0.002，...）。此外，在数据挖掘中，分析收集的数据以便在预测模型中使用它们之前突出显示主要信息。与其反对数据挖掘和统计数据，最好假定数据挖掘是致力于开发大型数据库的统计分支。所使用的技术来自不同领域，取决于经典统计和人工智能[29]。最后一个概念被定义为“构建计算机程序，这些计算机程序目前由人表现的更好，因为它们需要高级别的心理过程，如感知学习型组织记忆和批判性思维”。这个定义并没有真正的共识，还有很多类似的定义。

2.2.监督学习

在监督式学习中，计算机会给出一个由“教师”给出的示例输入及其所需输出，目标是学习将输入映射到输出的一般规则[23]。这些方法需要“专家”干预。训练数据包含一组训练示例。在监督学习中，每个模式都是一对，它包含一个输入对象和一个期望的输出值。监督式学习算法的功能是分析训练数据并产生推断的功能。

2.2.1。线性回归

研究时间序列的早期尝试，特别是在19世纪，通常以确定性世界的概念为特征。这是Yule（1927）的主要贡献，它通过假设每个时间序列可以被认为是随机过程的实现来发起时间序列随机性的概念。基于这个简单的想法，自那时以来已经开发了许多时间序列方法。 Slutsky，Walker，Yaglom和Yule等工作者首先提出了自回归（AR）和移动平均（MA）模型的概念[30]。沃尔德的分解定理[31]导致了Kolmogorov的线性预测问题在1941年的制定和解决。从那以后，在时间序列领域发表了大量的文献，涉及参数估计，识别，模型检验和预测;例如参见参考文献[32]进行早期调查。

2.2.2。广义线性模型

统计学中的广义线性模型（GLM）是普通线性回归的灵活泛化，它允许具有非正态分布的误差分布模型的响应变量。 GLM通过允许线性模型通过链接函数允许与响应变量相关并且将每个测量的方差的大小考虑为其预测值的函数来概括线性回归[33]。有些研究利用与其他预测因子（如卡尔曼滤波器）的耦合来提高回归质量[34]。

2.2.3。非线性回归

人工神经网络（ANN）越来越多地用于气象学中的非线性回归和分类问题，这是因为它们在数据分析和预测中的有用性[35]。在非线性方法的时间序列预测领域，ANN的使用尤其占优势。事实上，气象效用数据库的历史数据的可用性以及ANNs是能够在输入和输出变量集合之间执行非线性映射的数据驱动方法的事实使得该建模软件工具非常有吸引力。

具有d输入，m个隐含神经元和单个线性输出单元的人工神经网络定义了从输入向量x到输出y的非线性参数化映射（由无偏差形式给出）：

//公式编辑

每个m个隐藏单元通常与切线双曲线函数有关：。参数矢量管理非线性映射，并在一个称为训练或学习阶段的阶段进行估计。 在此阶段，人工神经网络使用包含一组n个输入和输出示例的数据集D进行训练。 第二阶段称为泛化阶段，包括在测试数据集D上评估人工神经网络的泛化能力，即在训练阶段没有看到例子时，给出正确的输出结果。

对于太阳辐射来说，输出和输入之间的关系的形式如下：

如前述方程所示，ANN模型等价于时间序列预测问题的非线性自回归（AR）模型。以与AR模型类似的方式，过去的输入值p的数量可以用自相互信息因子[36]来计算。

必须认真关注模型的构建，因为太复杂的人工神经网络很容易过度训练数据。人工神经网络的复杂性与隐藏单元的数量或相反矢量w的维数有关。可以采用修剪或贝叶斯正则化等几种技术来控制ANN的复杂性。经常使用Levenberg-Marquardt（逼近牛顿法）学习算法，在停止训练之前使用最大失败参数来估计ANN模型的参数。最大失败参数对应于一个正规化工具，用于限制特征数量的预测失败后的学习步骤，因此是控制模型复杂性的手段[18,37]。请注意，混合方法（如通过共轭梯度进行主控优化以选择ANN拓扑结构）可使ANN以最大容量运行。大量的研究表明人工神经网络与其他工具（如卡尔曼滤波器或模糊逻辑）之间的耦合的影响，这证实了增益常常是非常有趣的[38]。

2.2.4。支持向量机或支持向量回归

支持向量机是Vapnik在1986年引入的分类任务和回归问题中使用的另一种基于内核的机器学习技术[39]。支持向量回归（SVR）是基于支持向量机对回归问题的应用[18]。该方法已成功应用于时间序列预测任务。以与高斯过程（GP）类似的方式，由SVR机器为输入测试案例x计算的预测由下式给出：

通常利用RBF内核，如下定义：



参数b（或偏差参数）是从先前的方程和一些特定条件推导出来的。在SVR的情况下，系数αi与两个拉格朗日乘子的差异有关，这是二次规划（QP）问题的解。与面临局部最小问题的人工神经网络不同，这里的问题是严格凸的，QP问题有一个独特的解决方案。此外，必须强调的是（与全科医生不同），并非所有的培训模式都参与了前置关系。事实上，在QP问题中方便地选择成本函数（Vapnik’sεinsentive function）可以获得一个稀疏解。后者意味着只有一些系数ai将是非零的。带有非消失系数的例子称为支持向量。

在预测问题中使用SVR的一种方式与以下事实有关：给定训练数据集和一个测试输入向量x，预测的晴空指数可以计算出一个特定的地平线h，如：

2.2.5。决策树学习（Breiman装袋）

基本想法很简单。需要预测来自输入X1，X2，…，Xp的响应或类别Y。这是通过增长一棵二叉树来完成的。在树中的每个节点上，对其中一个输入进行测试，例如应用Xi。根据测试的结果，选择树的左侧或右侧子分支。最终到达一个叶节点，进行预测。该预测汇总或平均到达叶的所有训练数据点。通过使用每个独立变量来获得模型。对于每个单独的变量，均方误差用于确定最佳分割。每个分割中要考虑的最大特征数设置为特征总数[40-42]。

2.2.6。最近邻

最近邻居神经网络（k-NN）是一种基于实例的学习，其中函数仅在本地近似，所有计算都被延迟直到分类[37]。 k-NN算法是最简单的机器学习算法之一。对于分类和回归，分配一个权重给邻居的贡献是有用的，这样最近的邻居对平均值的贡献比远处的贡献更大。例如，在一个通用的权重安排中，每个邻居被赋予1 / d的权重，其中d是到邻居的距离[43]。

2.2.7。马尔可夫链

在预测领域，一些作者试图使用所谓的马尔可夫过程，特别是马尔可夫链。马可夫过程是具有马尔科夫性质的随机过程，这意味着在目前的状态下，未来的状态独立于过去的状态[44]。换句话说，对当前状态的描述完整地反映了所有可能影响过程未来发展的信息。在这个过程中，未来的状态是通过概率过程而不是确定性过程来达到的。这些过程的正确使用需要首先计算过渡态的矩阵。状态i到状态j的转移概率由pi，j定义。这些数字的家族称为马尔可夫链R的转换矩阵[27]。

2.3。无监督学习

与监督学习模型相反，无监督学习模型不需要“专家”干预，并且该模型能够在不知道输出的情况下在其输入中找到隐藏结构[45]。无监督学习与统计学密度估计问题类似。然而，无监督学习还包含许多其他技术，旨在总结和解释数据的关键特征。无监督学习中常用的许多方法都是基于用于预处理数据的数据挖掘方法。

2.3.1。 K-means和k-methods聚类

k均值聚类是一种矢量量化方法，最初是从信号处理中得到的，这种方法在数据挖掘中的聚类分析中很流行。 k均值聚类旨在将n个观测分为k个聚类，其中每个观测属于具有最近均值的聚类，用作聚类的原型。 k-Means算法专注于从数据中提取有用的信息，目的是对时间序列行为建模，并通过对数据进行聚类来查找输入空间的模式。此外，非线性自回归（NAR）神经网络是建模和预测非线性时间序列的强大计算模型[46]。有很多聚类方法可用，感兴趣的读者可以看到参考文献[47]了解更多信息。

2.3.2。分层聚类

在数据挖掘和统计中，层次聚类（也称为层次聚类分析）是一种聚类分析方法，旨在构建聚类的层次结构。分层聚类创建聚类的层次结构，其可以用称为“树状图”的树状结构来表示，其包括根和叶。树的根由一个包含所有观测值的单个簇组成，而树叶则对应于单独的观测值。层次聚类的算法通常是凝聚的，其中过程从树叶开始并连续地将聚类合并在一起或分裂，其中过程从根开始并递归地分裂簇[48]。任何不具有负值的函数都可以用作观察对之间相似度的度量。选择合并或分裂的聚类是由一个连接标准决定的，该连接标准是观测之间成对距离的函数。应该注意的是，在给定的高度切割树会以选定的精度给出聚类。

2.3.3。高斯混合模型

高斯过程（GPs）是非线性建模中的一个相对较新的发展[49]。 GP是将多元高斯分布推广到无限多变量的泛化。多元高斯分布d完全由平均向量m和协方差矩阵S来规定，例如，。GP建模的关键假设是数据可以表示为来自多变量高斯分布的样本，例如观察。为了更好地介绍GP，这种情况常常局限于一个标量输入变量x。由于数据通常是由测量误差产生的噪声，所以每个观测值y都可以被认为是一个基本函数f(x)，其中加入了独立的高斯噪声，方差为，即。由于GP是多变量高斯分布的一个扩展，它由一个均值函数m(x)和一个协方差函数充分说明。方程f(x)可以用不同的方式表达：。协方差函数的设置允许关联一个观察值yp到另一个yq。 协变函数的通常选择是。由于预测通常是使用噪声测量进行的，因此两个观测值之间的协方差可以表示为:

被称为克罗内克函数。 和l被称为超参数协方差函数，他们控制模型的复杂性，可以从手边的训练数据中学习（或优化）[49]。 例如，在预测研究中，给定训练数据集，对于新的测试输入X，对于预测期h的n预测辐射的矢量由预测高斯分布预测的均值给出。

2.3.4。集群评估

聚类中的典型目标函数正式形成了达到高群内相似性和低群集间相似性的目标[50]。这是聚类质量的内部标准。内部标准的良好分数并不一定意味着在应用程序中的良好效果。内部标准的一个替代方法是对利益申请的直接评估。对于搜索结果聚类，可能需要用户需要用不同聚类算法找到答案的时间量。这是最直接的评估，但是这非常耗时，特别是如果需要进行大量研究的话。

2.4。集成学习

集成学习的基本概念是将多个基础学习者作为集成成员进行训练，并将其预测结果合并为一个单一输出，与目标数据集上具有不相关误差的任何其他集成成员相比，其平均性能应该更好[51]。监督式学习算法通常被描述为执行通过假设空间搜索的任务来找到可以对特定问题执行良好预测的合适假设。即使假设空间包含对特定问题非常匹配的假设，也很难找到哪一个最好。集成将多个假设结合起来，创造出更好的假设。术语集合通常用于使用相同基础学习器生成多个假设的方法。诸如决策树之类的快速算法通常与集合一起使用，尽管较慢的算法也可以从集合技术中受益。与评估单个模型的预测精度相比，评估集成的预测准确性通常需要更多的计算时间，因此可以将集合考虑为通过执行更多计算来补偿差的学习算法的方式。多分类器系统的通用术语还涵盖了不是由相同的基础学习者引发的假设的结合。感兴趣的读者可以看到参考文献[52]关于集合学习的更多细节。

2.4.1。推进

集合模型使用决策树作为弱学习者，并通过优化损失函数以阶段式方式建立模型[34,35]。 Boosting作为一种将许多弱分类器结合起来以产生强大的“委员会”的方式出现。这是一个迭代过程，对坏的分类越来越重要。简单的策略可以显着提高分类性能。为此，使用所谓的弱学习者在递归线性预测的残差的每个时间范围上应用助推自回归过程，所谓弱学习者是具有相对于方差的大偏差的学习者。

2.4.2。套袋

Bootstrap聚合，也称为统计分类和回归中的bagging，是一种机器学习集成元算法，旨在提高机器学习算法的稳定性和准确性。该算法还可以减少方差并有助于防止过度拟合。虽然它通常应用于决策树方法，但它可以用于任何类型的学习方法。套袋是模型平均方法的特例。套袋预测指标通常用于生成预测变量的多个版本并使用它们来获得汇总预测变量。聚合在预测数值结果时平均所有版本，并进行多次投票以预测类别。多个版本是通过对学习集进行引导复制并将其用作新的学习集而形成的[55]。

2.4.3。随机子空间

在所提出的方法中使用的机器学习工具基于随机森林，该随机森林由多个决策树的集合或集合组成，每个决策树都是从一个样本中构建的，训练集，是一组输出。此外，在构建树时，只有一个随机的变量子集用于分割节点。因此，最终节点（或叶节点）可能包含一个或多个观察值。对于回归问题，每个树在呈现一组预测变量时能够产生响应，这是所产生叶上观测值的条件均值。条件均值通常用加权平均值近似。由于随机构建的结果，森林的偏倚通常稍微增加一个单一的非随机树的偏倚，但由于其平均值的方差减少，往往不仅仅是补偿的增加在偏见，从而产生一个整体更好的模型。最后，所有树木的响应也被平均以获得模型的单个响应变量，这里也使用加权平均值[56]。分类准确度的显着提高是由于种植一批树木并让它们投票选出最受欢迎的分类。为了生长这些集合体，通常生成管理集合体中每棵树生长的随机向量。首先使用的一个例子是装袋，其中为每棵树增长一个随机选择（没有替换）是由包含在训练集[57-60]中的例子构成的。

2.4.4。预测器集合

目前的实践表明，预测应该由许多简单的“常规”预测组成，或者根据其他简单预测（不仅包括点预测，还包括概率）进行简单预测。相对于贡献预测，这会带来性能上的提升。在统计模型的情况下，来自相同技术的实现（例如相同的神经网络体系结构）会多次进行训练，或使用不同的数据集样本;或不同的技术。一旦“第一阶段预测”可用，不同的组合方法是可能的。最简单的方法是对不同方法给出的结果进行平均。更通用的方法根据不同的标准和不同的权重策略，为每个时间范围的每个贡献方法分配一个权重。简单的预测可以被看作是对同一真实状态的不同理解。通过这种方式，不完美的传感器数据融合方法也应该是有效的，以执行预测的组合。基于集合的人工神经网络和其他机器学习技术已被用于全球辐射建模的许多研究中，并且与传统的回归模型相比，提供了更好的性能和泛化能力[28,42]。

3.评估模型的准确性

一般来说，评估是衡量某件事情有多好。这种评估用于模型开发的各个步骤，例如在预测模型本身的评估过程中（例如在统计模型的训练期间），用于判断经过一些修改后的模型的改进以及用于比较各种模型。如前所述，由于各种原因，例如不同的预测时间范围，预测数据的各种时间尺度以及从一个地点到另一个地点的气象条件的变化性等各种原因，这种性能比较都不容易。它通过将预测输出（或预测时间序列）与测量数据本身（或观测或测量时间序列）进行比较来工作，测量数据本身也与测量的误差（或精度）相关。

图形工具可用于通过实验测量来估计模型的适用性，例如：

- 预测辐照度的时间序列与测量的辐照度相比，可以很容易地看到预测的质量。在图4a中，作为一个例子，可以看到晴天情况下的高预测精度和局部多云情况下的低预测精度。

- 测量辐照度的预测散点图（见图4b中的例子），可根据辐照度条件揭示系统偏差和偏差，并显示与预测相关的偏差范围。

- 接受者操作特征（ROC）曲线比较真阳性和假阳性的比率。

没有采用标准的评估措施，这使得预测方法的比较困难。 Sperati等人[62]在欧洲行动可再生能源天气情报（WIRE）框架内提出了一项基准测试，目的是评估最新可再生能源预测模型的性能。这项研究是可靠性参数利用的一个很好的例子。他们的结论是：“为了全面了解所有可能的情况，需要进行更多的使用更多测试用例，数据和模型的工作。应该考虑遍布欧洲，美国和其他相关国家的测试案例，以努力代表大部分可能的气象条件“。本文很好地说明了性能比较的困难。

通常使用的统计信息包括以下内容：

平均偏差（MBE）代表预测的平均偏差：

通过预测输出（或预测时间序列），y观测数据（或观测或测量的时间序列）和N观测数量。 预测将低估或高估观测值。 因此，MBE并不能很好地指示模型的可靠性，因为这些误差相互补偿，但可以看出它高估或低估的程度。

平均绝对误差（MAE）适用于具有线性成本函数的应用，即由于预测不良导致的成本与预测误差成正比：

MSE通常是由训练算法最小化的参数。

均方根误差（RMSE）对较大的预测误差更为敏感，因此适用于容易出现小误差，更大误差导致成本过高的应用，例如实用应用[22]。 这可能是最受赞赏和使用的可靠性因素：

平均绝对误差百分比（MAPE）接近MAE，但观测数据和预测数据之间的差距除以观测数据以考虑相对差距：

该指数有一个缺点，即当y(i)接近于零时它不稳定，并且不能为y(i)=0定义。

通常，这些误差是RMSE特别标准化的; 作为参考，通常使用辐照的平均值，但可以找到其他定义：

y是y的平均值。 其他指数存在，可以用作相关系数R（皮尔森系数），或者一致指数（d），它们在0和1之间归一化。

由于预测准确性强烈依赖于评估所用的地点和时间以及其他因素，因此很难仅从精度指标评估预测质量。然后，最好将不同预测的准确度与一组通用测试数据进行比较[63]。 “微不足道的”预测方法可以用作参考[22]，最常见的是持续模型（“事物保持不变”，[64]），其中预测总是等于最后一个已知数据点。持久性模型在预测文献中也被称为幼稚模型或RandomWalk（一系列随机步骤组成的路径的数学形式化）。由于太阳的几何路径，太阳辐照度具有确定性分量。作为一个例子，这个特性可以作为一个约束条件被添加到最简单的持续形式中，前一天的测量值或前一个小时同时作为预测值。其他常见的参考预测包括基于气候常数和简单自回归方法的预测。这种与参考NWP模型的

比较如图5所示。一般来说，1小时后的预测好于持续性。对于超过两天的预报视野，气候平均值显示较低的误差，应该是首选。

通常，使用参考模型进行性能比较，然后使用技能因子。技能因素或技能得分定义了预测与参考预测之间的差异，归一化了完美参考预测与参考预测之间的差异[18]：

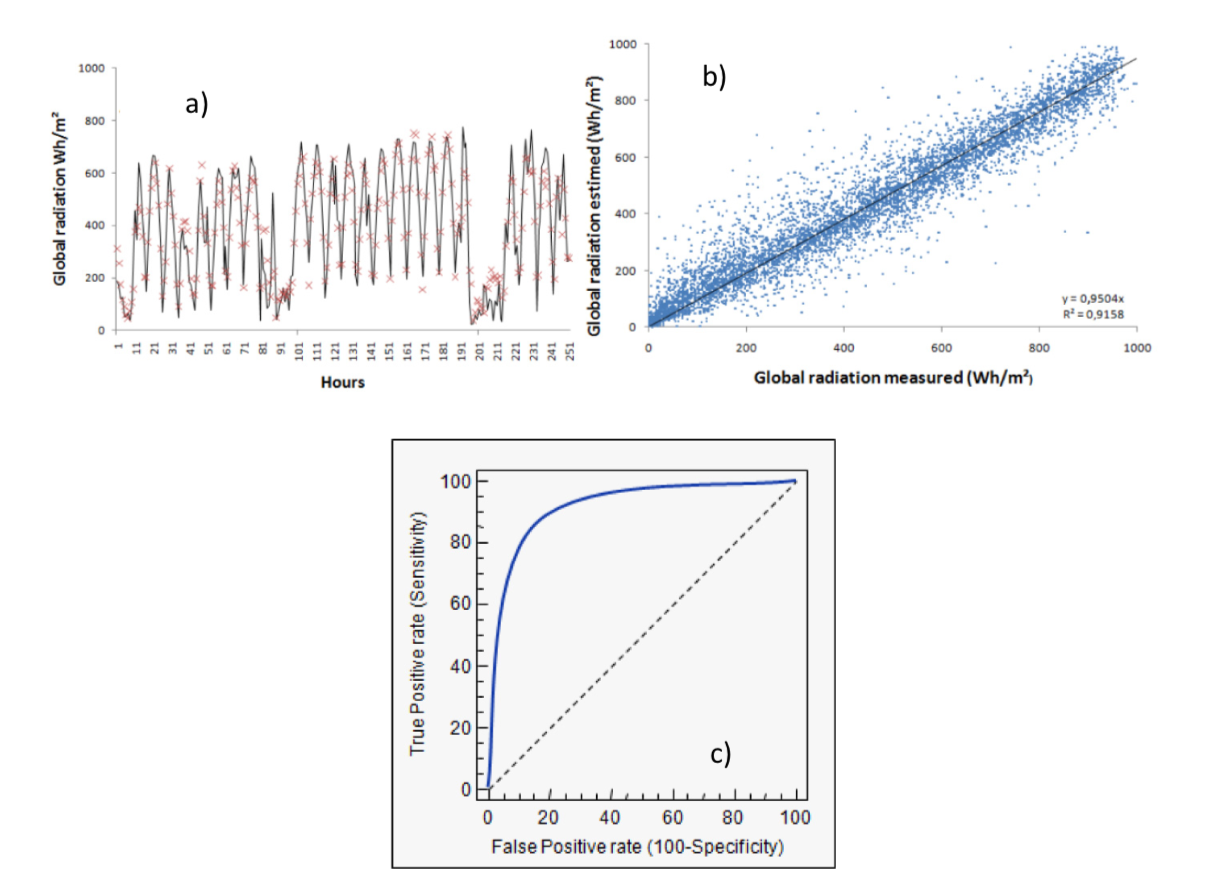
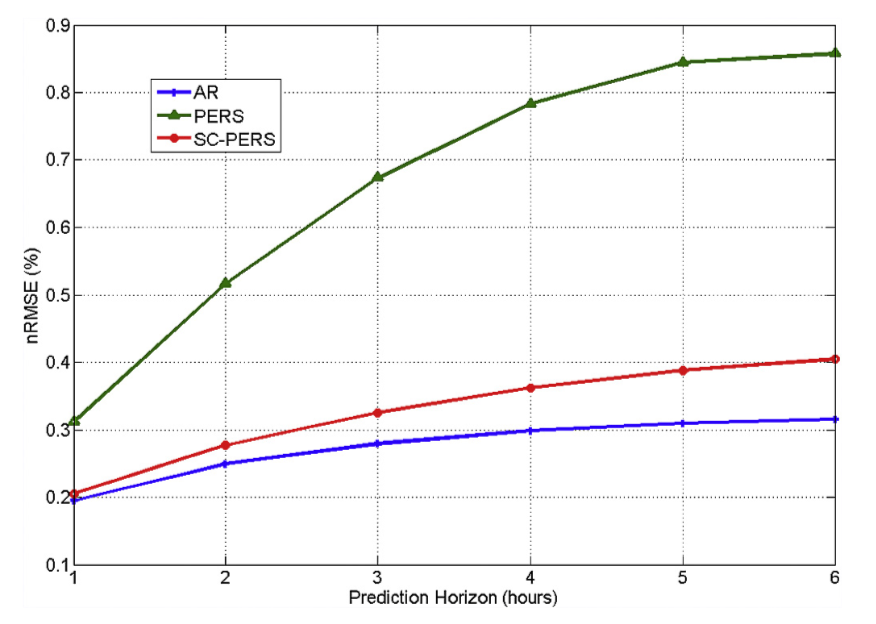
其值因此在1（完美预测）和0（参考预测）之间变化。负值表示比参考值更差的表现。技能分数不仅可用于与简单参考模型进行比较，而且还可用于不同预测方法（改进分数）的比较。作为一个例子，Bacher等人[65]报告RMSE相对于持续性的改善了36％，那么相对于持续性的RMSE技能评分等于0.36。标杆管理也可用于确定预测相对较好的条件。在美国[17]，加拿大和欧洲国家[8]和意大利[62]实现了许多基准测试。请注意，文献中的太阳能预测方法仅限于点预测。概率预测也被广泛使用，并且通常是更实际的太阳能需求解决方案。概率/预测间隔预测的评估是不同的，所使用的指标不限于所提出的预测间隔（参见使用预测间隔[66,67]）。

图4 a）2008年阿雅克肖（法国）预测和测量的全球辐照时间序列; b）阿雅克肖（法国）的预测与实测全球辐照度的散点图; c）ROC曲线的示例（理想的ROC曲线靠近左上角）。

图5 预测的相对均方根误差（持续性，自回归和标度持续性）以及取决于预测时限的参考模型[18]

4.机器学习预测者的比较

在展示与机器学习方法有关的结果以预测全球辐射之前，图6显示了在太阳能预测的5种主要期刊中，人工神经网络，机器学习和SVM / SVR被引用的次数（太阳能，能源，应用能源，可再生能源和能源转换和管理）。

这三个术语越来越多地用于文学。可以看出，ANN是全球辐射预测中最常用的方法。

4.1。 ANN案例

参考文献中介绍了这种预测方法的评论[28,47]感兴趣的读者可以在这些论文中找到更多信息。神经网络已经在世界许多地方进行了研究，研究人员已经显示出这些技术能够准确预测气象数据的时间序列。区分两种类型的研究至关重要:建模与多元回归和时间序列预测。事实上，MLP经常用于“通用逼近”的属性，能够进行非线性回归。 1999年，作者首次提出了通过MLP预测全球太阳辐射时间序列。 Kemmoku [69]使用基于MLP的方法来预测第二天的辐射。结果显示夏季预测误差（MAPE）为18.5％，冬季为21.8％。从与ANN [28,47]相关的所有文章中，与预测（月，日，小时和分钟）相关的误差在5％到15％之间。在Mellit和Kalogirou的综述文章[68]中，可以看到79％的天气预报数据中使用的人工智能（AI）方法基于联结主义方法（ANN）。使用模糊逻辑（5％），自适应神经模糊推理系统（ANFIS）占论文的5％，网络耦合小波分解和ANN为8％，混合ANN 和马尔可夫链为3％。总结使用人工神经网络，尤其是MLP代表了绝大多数的研究工作。这是最常用的技术。其他方法仅偶尔使用。根据已发表的文献，影响预测的参数很多，所以很难采用其他研究的结果。考虑到这个事实，即使它们未必在其他研究中得到证实，测试方法或参数也许是有趣的。根据上述内容，必须针对每个测试地点研究MLP或ARMA方法固有的所有参数。即使MLP看起来比ARMA更好，在少数情况下也会出现相反情况。正是在这种背景下研究了其他机器学习方法。请注意，统计指标（RMSE，预测技能）方面的预测性能不仅取决于天气条件（变化性），还取决于预测范围。

4.2。单机学习方法

表3显示了用于全球太阳辐射预测，直接正常辐照度和漫射辐照度的机器学习方法的结果。这里没有引用很多论文。感兴趣的读者可以看到一些与这个主题相关的书面评论文章[26,47,49]。

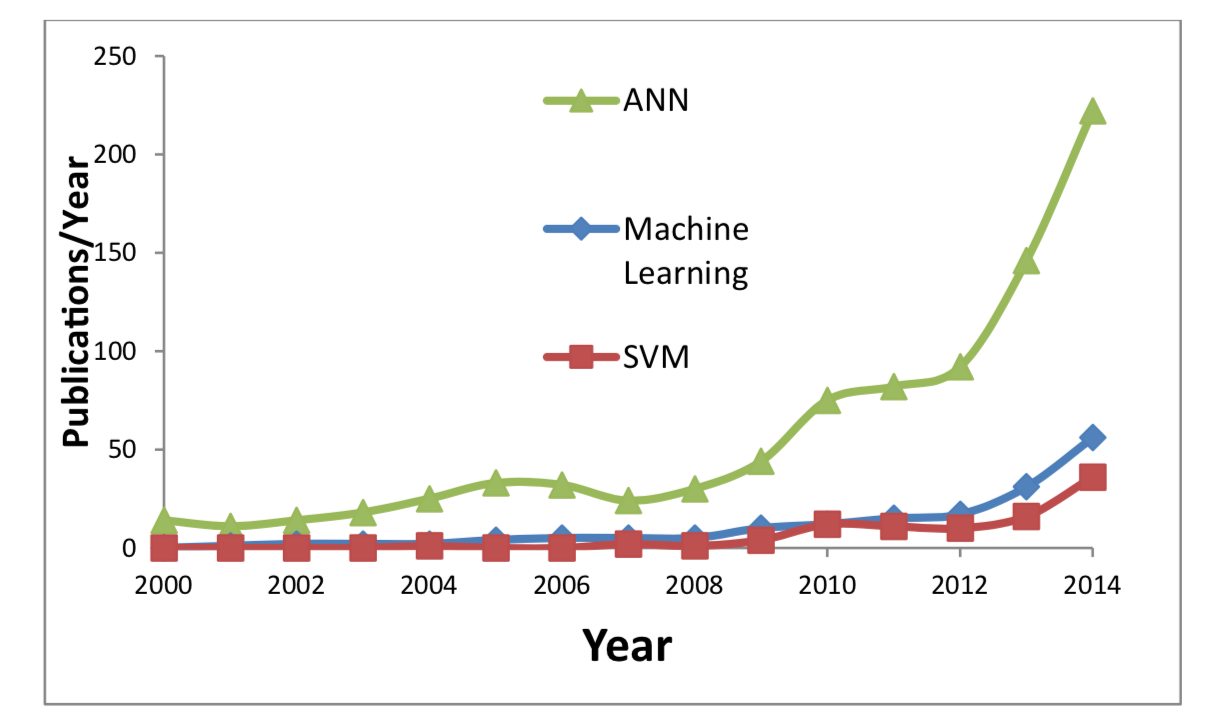
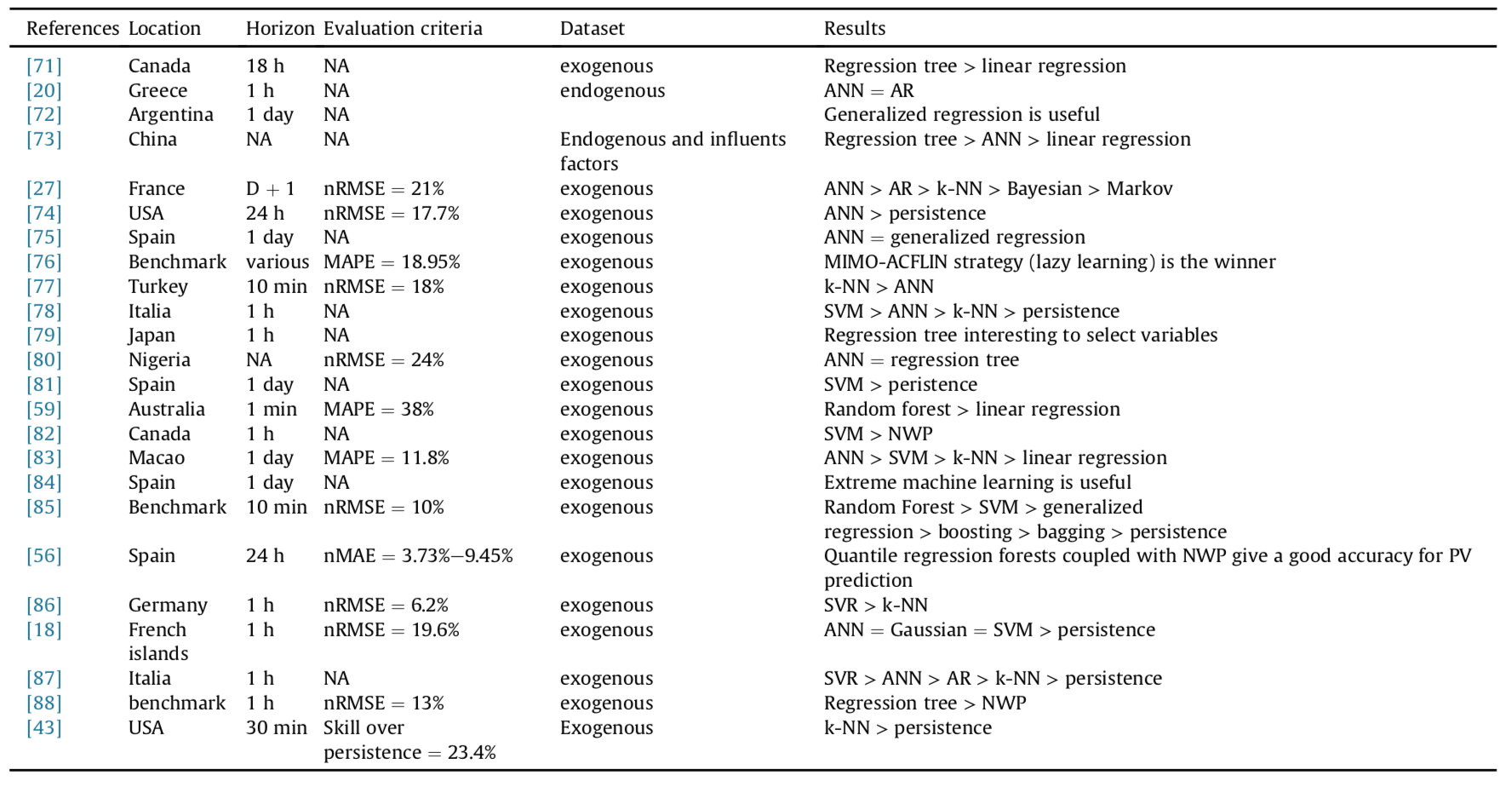
在这个列表中，可以看出，使用SVR，ANN，k-NN，回归树，提升，装袋或随机森林的论文比传统的回归方法在系统上获得了更好的结果。 ANN和SVM在预测方面给出了类似的结果，但可以得出结论：SVM比ANN更容易使用;优化步骤是自动的，但在ANN情况下非常复杂。因此，最好使用SVM而不是ANN。所有与使用回归树或类似方法有关的方法（助推，装袋或随机森林）很少使用，但会产生极好的结果。在这个阶段得出结论并不容易，但在未来5年，这些方法很可能成为辐照预测方面的参考。以前的所有方法的替代方案肯定是k-NN，但是需要更多的公布才能得出结论：它是一个好预报者。实际上，虽然SVR，回归树，提升，装袋和随机森林似乎是最有效的，但提出机器学习方法的排序是非常困难的。为了克服这个排名问题，一些作者毫不犹豫地将单个预测因子结合起来。

图 6 在原始文章中使用了ANN，机器学习和SVM术语的时间

表3 与使用单纯机器学习方法的全球太阳辐射预测有关的代表性论文清单

4.3。预测因子集合

如表4所示，有很多结合了预测因子的解决方案。在这些论文中，我们可以看到，经常使用ANN来构建预测因子集合（> 70％的情况）。

正如可以看到的那样，系统地预测因子集合比单个预测因子有更好的结果，但是最好的杂交方法并没有真正定义。为了提出一个可靠的方法，或者为了证明所有的方法是等价的，需要做更多的工作，但当然这不能用这里提出的有限的情况来完成。

5.结论和展望

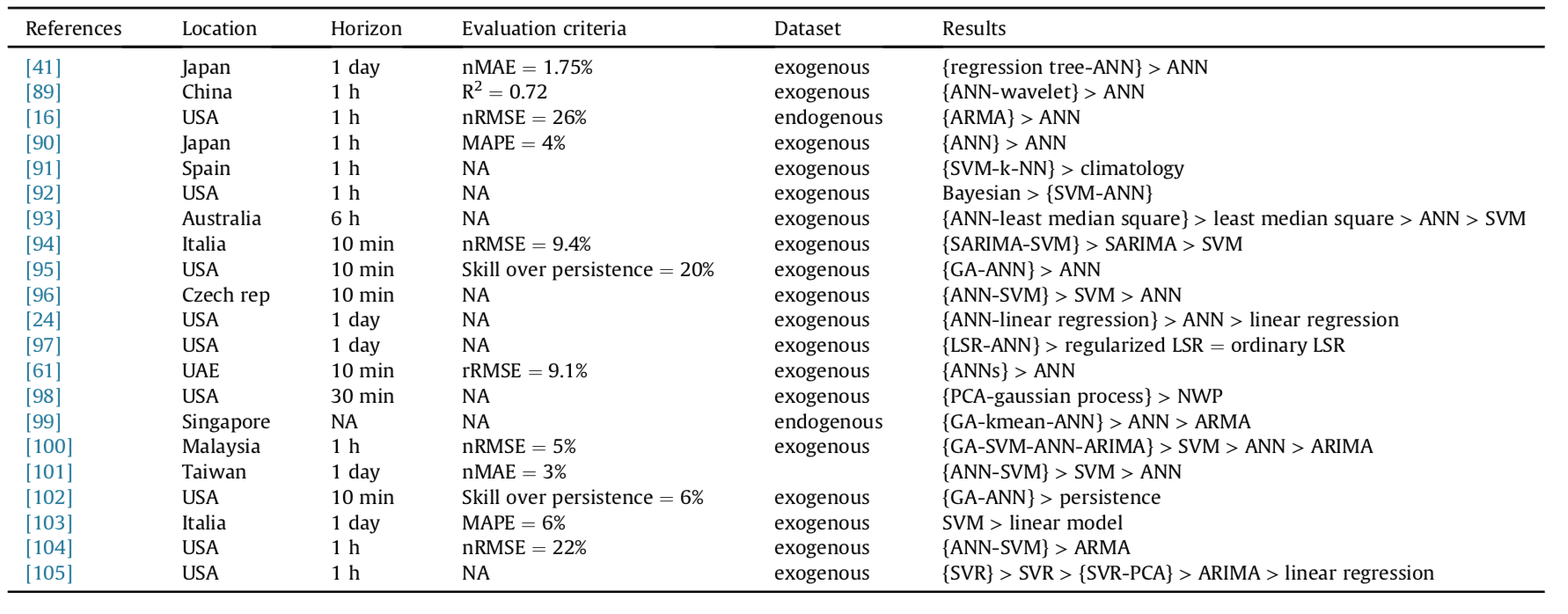
如本文所示，许多方法和类型的方法都是可用的。有太多的方法来估计太阳辐射，有些常常被使用（ANN，ARIMA，天真的方法），其他的开始被更频繁地使用（SVM，SVR，k-均值），其他很少被使用（增强，回归树，随机森林等）。在某些情况下，一个是最好的，而另一个则是相反的。作为结论，可以说ANN和ARIMA方法在某些变化条件下的预测质量方面是等价的，但ANN作为通用非线性逼近的灵活性使得它们比经典ARIMA更可取。通常，这些方法的准确性取决于训练数据的质量。未来几年应该普遍使用的三种方法是支持向量机，回归树和随机森林，因为所给出的结果非常有前景，并且未来几年肯定会产生一些有趣的研究。实际上，考虑到已发表的论文，这些方法产生了类似的错误统计。这些方法的实施可能与文献报道的错误有关，而不是方法本身。例如，当相同输入的误差自相关降低白噪声时，SVM，SVR，回归树或随机森林表现非常相似，两者之间没有统计学差异。从表4可以看出，第二点是事实上，预测器集成方法总比单纯的预测器好。这显示了必须研究问题的方式，而仅使用一种随机方法（首先是ANN和ARIMA）的简单预测方法应该趋于消失。在本文中，深度学习是机器学习的一个分支，它基于一组试图使用模型架构对数据进行高级抽象建模的机器学习分支，具有复杂结构或由多个非线性变换组成的复杂结构，没有考虑到。这个研究领域是非常近的，并且没有足够的经验，但是在未来，这种方法可能会胜过传统方法，就像其他预测领域（空气质量，风力，经济等）的情况一样。因此，为了满足各种需求，可以计算通过各种方法达到的预测结果。那么问题就出现了，他们将如何组合起来。答案显然不是微不足道的，因为各种结果预测在许多方面显示出不同。而且，其中一些将与置信区间相关联，这也应该合并。

表4 有关全球辐射预测与机器学习方法相结合的代表性论文清单

答谢

这项工作是在地平线2020项目（H2020-LCE-2014-3 - LCE-08/2014 - 646529）TILOS“地方规模的技术创新，电池储能的最佳整合”框架内实施的。 作者想感谢欧盟的财政支持。