第 42 卷 第 10 期
2020 年 10 月

海 洋 学 报
Haiyang Xuebao

Vol. 42 No. 10
October 2020

钟国荣, 李学刚, 曲宝晓, 等. 基于广义回归神经网络的全球表层海水 1°×1°二氧化碳分压数据推演[J]. 海洋学报, 2020, 42(10): 70–79, doi:10.3969/j.issn.0253−4193.2020.10.007

Zhong Guorong, Li Xuegang, Qu Baoxiao, et al. A general regression neural network approach to reconstruct global 1°×1° resolution sea surface pCO_2[J]. Haiyang Xuebao, 2020, 42(10): 70–79, doi:10.3969/j.issn.0253−4193.2020.10.007

基于广义回归神经网络的全球表层海水 1°×1°二氧化碳分压数据推演

钟国荣[1,2,3,4], 李学刚[1,2,3,4*], 曲宝晓[1,3,4], 王彦俊[4], 袁华茂[1,2,3,4], 宋金明[1,2,3]

(1. 中国科学院海洋研究所 海洋生态与环境科学重点实验室, 山东 青岛 266071; 2. 中国科学院大学, 北京 100049; 3. 青岛海洋科学与技术试点国家实验室 海洋生态与环境科学功能实验室, 山东 青岛 266237; 4. 中国科学院 海洋大科学研究中心, 山东 青岛 266071)

摘要: 表层海水二氧化碳分压是评估海洋碳源汇强度的关键参数, 但其实测数据较少、时空分布极不均匀, 导致二氧化碳交换通量的估算有很大的不确定性, 海洋源汇特征就不能确切获取。为了解决这个难题, 在收集的表层大洋二氧化碳地图 (Surface Ocean CO_2 Atlas, SOCAT) 实测数据集基础上, 运用广义回归神经网络建立二氧化碳分压与经纬度、时间、温度、盐度和叶绿素浓度间的非线性关系, 构建了 1998–2018 年间全球 1°×1°经纬度的表层海水二氧化碳分压格点数据, 其标准误差为 16.93 μatm, 平均相对误差为 2.97%, 优于现有研究中的前反馈神经网络、自组织映射神经网络和机器学习算法等方法。根据构建的数据所绘制的全球表层海水二氧化碳分压的分布与现有研究有较好的一致性。

关键词: 广义回归神经网络; 表层海水二氧化碳分压; 全球大洋格点数据

中图分类号: P714.2$^+$2 **文献标志码:** A **文章编号:** 0253−4193(2020)10−0070−10

1 引言

当前的研究普遍认为大洋每年可以吸收约 2 Pg (以碳计)左右的大气 CO_2, 这一结果主要是通过海−气二氧化碳分压差估算出来的, 并且与模式的估计也相一致。但用分压差估算出的结果仍有很大的不确定性, 主要原因是二氧化碳海−气交换速率的不确定, 以及参与计算的表层海水二氧化碳分压(pCO_2)的数据较少且空间分布不均匀[1−3]。尽管 pCO_2 实测数据相对一些其他参数比较容易获得, 可以通过基于非色散红外法的船舶连续走航观测获得[4], 但获得的实测数据相对于整个大洋来说仍然较少, 这使得通过集成多年的数据构建气候态分布成为过去比较有效的研究方法[5], 但要获得大范围区域 pCO_2 的连续时间变化用一般的空间插值法仅依靠这些实测数据是远远不够的。并且实测数据在时空分布上也非常不均匀, 特别是早期 20 世纪 70 年代前的实测数据几乎没有, 这极大地限制了基于 pCO_2 演化有关的大洋碳循环研究的时空尺度。虽然美国国家航空航天局 (NASA)、欧洲多国合作观测项目(EPOCA)等一直在着手扩展海洋观测网络, 但巨大的资金投入换来的也是十分有限的时空覆盖范围。在这个背景下, 通过大数据技术利用仅有的少量观测数据和一些辅助参数, 构建均匀的大洋 pCO_2 格点数据来研究全球变化成为新的突破方向。有研究者尝试利用传统的多元线性回归来重建二氧化碳分压变化, 但其结果只适用于有限的特定区

收稿日期: 2019−12−29; **修订日期:** 2020−03−23。

基金项目: 国家重点研发计划(2017YFA0603204); 国家自然科学基金(91958103); 中国科学院战略性先导科技专项(XDA19060401)。

作者简介: 钟国荣(1996—), 男, 江西省赣州市人, 研究方向为大数据技术在海洋化学中的应用。E-mail: zhongguorong18@mails.ucas.ac.cn

*** 通信作者:** 李学刚, 男, 研究员, 主要从事海洋生物地球化学方面的研究。E-mail: lixuegang@qdio.ac.cn

域[6], 甚至只适用于特定季节[7]。相比之下机器学习算法和人工神经网络更具有优势, 可以通过实验建立起大量参数间的实证关联, 来更准确地反映复杂的海水系统中 pCO_2 的变化规律[8]。机器学习算法包括随机森林算法(Random Forest Algorithm, RFRE)、支持向量机(Support Vector Machine, SVM)等, 目前的研究也多局限于单一过程主导的小范围区域, 对复杂区域及全球范围的预测则显得比较乏力。人工神经网络种类很多, 现有研究中利用的有前反馈神经网络(Feed forward Neural Network, FFNN)[9−10]、自组织映射神经网络(Self-Organizing Map, SOM)[11] 等, 目前仍存在较大的不确定性, 其标准误差从 17.6 μatm 到 20.2 μatm 不等[12−13]。广义回归神经网络(General Regression Neural Network, GRNN)是 FFNN 中径向基网络的一种变形形式, 与传统的前反馈网络相比, GRNN 是非线性拟合能力特化的形式, 在各个学科和工程领域应用都更加广泛。GRNN 无需传统的改变神经元间连接权值的训练, 只需要对一个光滑因子寻优, 训练速度比 FFNN 快几十到上百倍, 对数据预测的连续性也优于 SOM 的离散估计。为了获得误差更低的高时空分辨率全球表层海水 pCO_2 数据, 本文首次尝试了将 GRNN 应用于表层海水 pCO_2 格点数据的推演。

2　数据来源

研究使用的表层海水二氧化碳分压实测数据来源于表层大洋二氧化碳地图(Surface Ocean CO_2 Atlas, SOCATv2019)数据集, 该数据集由超过 100 个成员组成的国际海洋碳研究组织组建, 对实测数据进行了质量控制后公开发布。整个数据集包含约 2 570 万条观测数据, 时间范围为从 1957−2018 年。由于受叶绿素浓度数据的时间范围限制, 我们只使用了 1998−2018 年的数据。数据的总数量分布如图 1 所示。实测数据的分布十分不均匀, 整体上北半球数据覆盖率和数据总量都高于南半球, 欧洲、日本与美国东部沿岸等少数区域数据总量超过 10 万条, 而印度洋、南太平洋和一些近岸区域 20 年间数据总量只有 100 到 1 000 条左右, 不到 10 条甚至没有数据的区域也占不小的比例。在时间分布上不均匀的程度更加明显, 如图 2 所示的最近 20 a 获得的 pCO_2 调查数据, 以后 10 a 数据量多, 数据覆盖范围更广, 而前 10 a 数据量少, 覆盖范围也小。

图 1 图例中 n 为数据的数量级, 代表格点位置中有 10^n 条实测数据, 灰色部分代表格点位置中无实测数据。

图中实测数据覆盖范围指有实测数据的网格数占海洋区域总网格数的比例, 1°×1°经纬度的分辨率下, 海洋区域总网格数约为 43 000 个。

在 SOCAT 数据集中给出的是二氧化碳逸度(fCO_2), 在使用时将其转换成二氧化碳分压, 以便与其他研究或数据集进行对比验证, 二者间的换算关系为[14]

$$fCO_2 = pCO_2 \cdot \exp\left(p \cdot \frac{B+2\delta}{RT}\right), \tag{1}$$

式中, R 为气体常数(8.314 J/(K·mol)); p 为大气压(单位: Pa); T 为绝对温度(单位: K); B 和 δ 为校正系数(单位: m³/mol), 计算式为

$$B = \left(-1\,636.75 + 12.040\,8T - 3.279\,57\times10^{-2}T^2 + 3.165\,28\times10^{-5}T^3\right)\times10^{-6}, \tag{2}$$

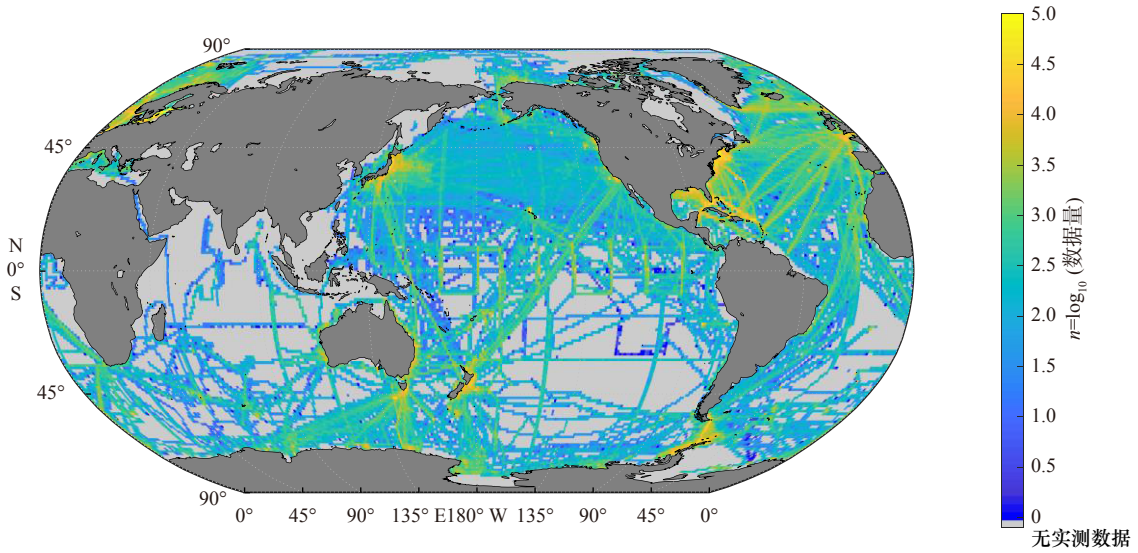


图 1　1998–2018 年间 SOCAT 二氧化碳分压数据分布情况

Fig. 1　Spatial distribution of pCO_2 observations in SOCAT dataset from 1998 to 2018

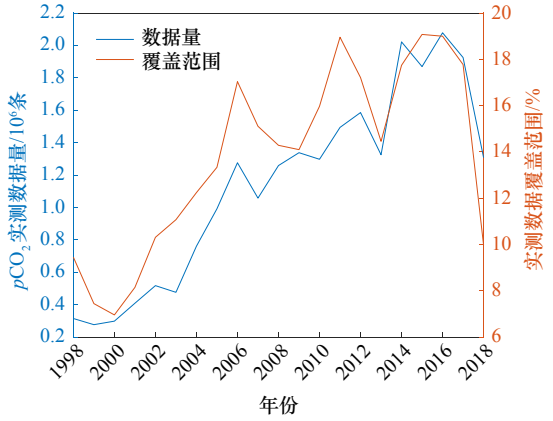


图 2 SOCAT pCO_2 实测数据时间分布

Fig. 2 Temporal distribution of pCO_2 observations in SOCAT dataset

$$\delta = (57.7 - 0.118T) \times 10^{-6}. \tag{3}$$

理论上，表层海水二氧化碳分压主要受海水的热力学性质、生物活动和物理过程控制。在新构建的方法中，选取了与这些过程紧密相关的温度、盐度和叶绿素浓度，加上与时空连续性相关的经纬度、时间等参数作为推演 pCO_2 的辅助参数。这些辅助参数的实测数据时空覆盖范围决定了生产出的格点数据的时空覆盖范围，因此在相关性高的条件下，应优先选择实测数据多的参数。本方法中使用的表层海水温度与叶绿素浓度为卫星遥感数据，具有足够大的空间范围和足够长时间的连续观测。通过建立与这些参数间的非线性关系来推演表层海水二氧化碳分压变化：

$$pCO_2 = f(Lon, Lat, Year, Month, SST, SSS, CHL), \tag{4}$$

式中，*Lon*、*Lat* 为经过三角函数换算的经纬度，经度为 0°～360°制，以保证数据在空间上的连续性。*Year*、*Month* 分别为数据对应的年和月；*SST*、*SSS* 分别为表层海水的温度（单位：℃）、盐度；*CHL* 为叶绿素浓度（单位：mg/m³）。使用的所有参数实测数据来源如表 1 所示。

$$Lon = \sin\left(\frac{lon}{360} \times \pi\right), \tag{5}$$

$$Lat = \cos\left(\frac{lat}{180} \times \pi\right). \tag{6}$$

3 格点数据构建方法

3.1 广义回归神经网络原理

广义回归网络是 Specht[15] 在 1991 年建立的一种径向基网络的变形形式，和径向基网络一样具有良好的非线性问题处理能力，并且训练更为方便。将训练样本作为后验条件，在 Parzen 非参数估计的基础上，广义回归网络计算输出时遵循最大概率原则[16]。

假设神经网络的输入和输出分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，联合概率密度可表示为 $f(\boldsymbol{X}, \boldsymbol{Y})$，以 $\boldsymbol{X}_0$ 代表训练集的观测值输入，$\boldsymbol{Y}$ 相对 $\boldsymbol{X}$ 的回归为

$$E(\boldsymbol{Y}|\boldsymbol{X}_0) = (\boldsymbol{X}_0) = \frac{\int_{-\infty}^{0} \boldsymbol{Y} f(\boldsymbol{X}_0, \boldsymbol{Y}) \mathrm{d}\boldsymbol{Y}}{\int_{-\infty}^{0} f(\boldsymbol{X}_0, \boldsymbol{Y}) \mathrm{d}\boldsymbol{Y}}. \tag{7}$$

输入 $\boldsymbol{X}_0$ 时，神经网络的预测输出为 $\boldsymbol{Y}(\boldsymbol{X}_0)$。给出训练样本数据集 $\boldsymbol{X}_0$ 与 $\boldsymbol{Y}_0$ 的情况下，利用 Parzen 非参数估计对密度函数 $f(\boldsymbol{X}_0, \boldsymbol{Y})$ 进行估算并化简可以得到：

$$\boldsymbol{Y}(\boldsymbol{X}_0) = \frac{\sum_{i=1}^{n} \left[\mathrm{e}^{-\mathrm{d}(\boldsymbol{X}_0, \boldsymbol{X}_i)} \int_{-\infty}^{+\infty} \boldsymbol{Y} \mathrm{e}^{-\mathrm{d}(\boldsymbol{Y}_0, \boldsymbol{Y}_i)} \mathrm{d}\boldsymbol{Y} \right]}{\sum_{i=1}^{n} \left[\mathrm{e}^{-\mathrm{d}(\boldsymbol{X}_0, \boldsymbol{X}_i)} \int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{d}(\boldsymbol{Y}_0, \boldsymbol{Y}_i)} \mathrm{d}\boldsymbol{Y} \right]}, \tag{8}$$

$$\mathrm{d}(\boldsymbol{X}_0, \boldsymbol{X}_i) = \sum_{j=1}^{l} \left[\left(\boldsymbol{X}_{0j} - \boldsymbol{X}_{ij} \right) / \sigma \right]^2, \tag{9}$$

式中，n 为样本总数，l 为输入变量 $\boldsymbol{X}$ 的维数。σ 为光滑因子，等同于高斯函数中的标准差。$\boldsymbol{X}_i$ 代表第 i 个计算样本对应的神经网络输入，$\boldsymbol{Y}_i$ 代表第 i 个计算样本对应的神经网络输出；$\boldsymbol{X}_{0j}$ 代表训练样本数据集输入 $\boldsymbol{X}_0$ 的第 j 个维度，$\boldsymbol{X}_{ij}$ 代表第 i 个计算样本对应的输入 $\boldsymbol{X}_i$ 的第 j 个维度。

由于 $\int_{-\infty}^{+\infty} \boldsymbol{Y} \mathrm{e}^{-Y^2} \mathrm{d}\boldsymbol{Y} = 0$，由式（8）化简得到：

$$Y(\boldsymbol{X}_0) = \frac{\sum_{i=1}^{n} \boldsymbol{Y}_i \mathrm{e}^{-\mathrm{d}(\boldsymbol{X}_0, \boldsymbol{X}_i)}}{\sum_{i=1}^{n} \mathrm{e}^{-\mathrm{d}(\boldsymbol{X}_0, \boldsymbol{X}_i)}}. \tag{10}$$

表 1 数据来源

Table 1 Data source

参数	时间范围	分辨率	数据来源
pCO_2	1957–2018年	–	Surface Ocean CO_2 Atlas (https://www.socat.info/)
盐度	1940–2018年	1°×1°	IAP, Global Ocean Heat Content Change (http://159.226.119.60/cheng/)
叶绿素	1997–2019年	1°×1°	European Service for Ocean Colour, Globcolour Project (http://www.globcolour.info/)
温度	1981–2019年	0.25°×0.25°	NOAA OI SST V2 High Resolution Dataset (https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.highres.html)

式(10)即为广义回归神经网络计算出的预测值, 其分子为训练集中所有样本求出的 $\boldsymbol{Y}_i$ 的加权和, 权值等于 $e^{-d(\boldsymbol{X}_0,\boldsymbol{X}_i)}$。

从结构上看, 广义神经网络分为 4 层: 输入层、隐含层、加和层和输出层 (图 3)。

输入层即负责接收样本数据的输入向量 $\boldsymbol{X}$, 神经元数量与输入向量 $\boldsymbol{X}$ 的维数 l 相同, 以简单的线性函数作为传输函数。其中维数 l 等于 7, 即输入包含 7 个维度, 分别为经度、纬度、年、月、海表温度、海表盐度和叶绿素浓度。一些研究中时间仅使用月或者仅使用年作为辅助参数, 但同时使用能略微降低整体的误差, 因为增加了输入向量的维度, 而且这对计算时间影响很小。隐含层的神经元数量为驯良样本数量, 通常使用高斯函数作为基础函数, Φ_i 代表第 i 个隐含层神经元, 其中心向量为对应的输入向量 $\boldsymbol{X}_i$。加和层的神经元只有两种, 分别为分子单元和分母单元。分子单元将训练集样本的输出期望作为权值, 求得隐含层神经元的加权和, 即式(12)中的分子部分, 分母单元负责的是隐含层神经元的代数和, 即式(12)中的分母部分, 分子单元和分母单元的输出在输出层中相除即得到输入 $\boldsymbol{X}$ 对应的预测输出 $\boldsymbol{Y}$。

3.2　网络训练与插值

为快速检索, 原始实测数据根据时间和经纬度存储在细胞数组中, 时间分辨率为 12 月×21 a, 空间分辨率为 1°×1°经纬度, 对神经网络进行训练时需要先将格点化的数据集转换成向量, 再将其输入到网络中, 过程如图 4 所示。

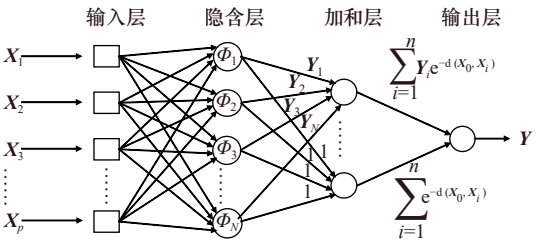


图 3　广义回归神经网络结构

Fig. 3　Structure of general regression neural network

训练使用的样本数据集为 1998−2018 年的所有数据中随机抽取的 80%, 由于数据总量大, 并且格点数据构建的目标最小时间分辨率为 1 个月, 对同 1 个 1°×1°格点里同 1 个月内的 pCO_2 实测数据进行了算术平均化处理。GRNN 程序通过 MATLAB 自带的广义回归神经网络函数工具箱实现, 网络的创建、训练和插值计算均可以通过工具箱函数命令进行。其中训练过程在创建的同时完成, 函数语法为

$$net = \mathrm{newgrnn}(\boldsymbol{X}_0, \boldsymbol{Y}_0, spread), \tag{11}$$

式中, $\boldsymbol{X}_0$、$\boldsymbol{Y}_0$ 分别为训练集的输入和对应的期望输出, $\boldsymbol{X}_0$ 是经度、纬度、年、月、温度、盐度、叶绿素浓度实测数据组成的向量, $\boldsymbol{Y}_0$ 是 pCO_2 实测数据组成的向量; *net* 为网络名, 在同时存在多个网络时用于辨识; newgrnn 为 MATLAB 自带的广义回归网络工具箱函数, 用于创建并训练网络。*spread* 为扩散速度, 是人为设定的固定标量, 默认为 1.0, 其值越大拟合出的曲线越平滑, 但如果想更精确地接近训练样本的期望输出, 应该选择较小的扩散速度值, 经过多次试验后我们择优选取的值为 1.4。广义回归网络的训练过程目的是为了求得式(9)中光滑因子 σ 值的最佳值, 这个值很大程度地影响网络的性能。当 σ 值非常大时, $d(\boldsymbol{X}_0, \boldsymbol{X}_i)$ 趋近于 0, 计算出的输出 $\boldsymbol{Y}(\boldsymbol{X}_0)$ 近似于所有训练集样本输出的平均值; 当 σ 值趋近于 0 时, 神经网络会出现过学习的现象, 表现为给定的输入与训练集中某一数据相同时, 计算得到的预测输出与实测值非常接近, 但给出的输入不在训练数据集中时, 计算出的输出与实测值偏差较大。避免这各种情况出现的方法是对输入的各个参数量级进行调整, 保证各参数

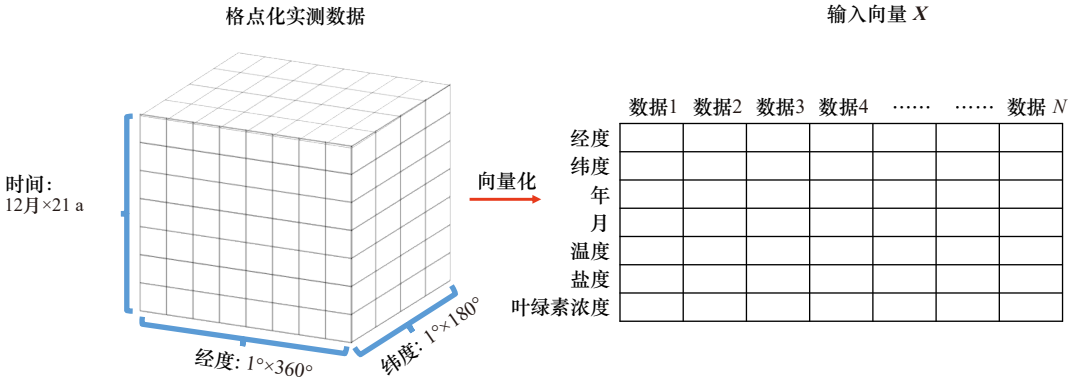


图 4　原始数据向量化过程

Fig. 4　Vectorization of original data

变化范围的数量级不相差过大。在我们使用的输入参数中，除年份外的参数均在 0～40 间变化，因此将所有数据的年份的数量级调整到与其他参数一致：

$$Year\left(\text{调整后}\right) = Year\left(\text{调整前}\right) - 1997. \quad (12)$$

调整后年份的变化范围为 1～21，这样就避免了过学习现象的出现。

创建并训练网络后，输入二氧化碳分压空白区域对应的温、盐等参数组成的向量 $\boldsymbol{X}$，即可计算出预测的二氧化碳分压值 $\boldsymbol{Y}$，函数语法为

$$\boldsymbol{Y} = net(\boldsymbol{X}), \quad (13)$$

式中，*net* 和创建时的 *net* 为网络名，可替换为其他名称，但两个过程的名称必须保证一致。在这个计算过程中，输入的向量 $\boldsymbol{X}$ 包括了所有待插值的格点，不需要每个格点单独计算。最后再将输出的二氧化碳分压预测值向量 $\boldsymbol{Y}$ 还原成 180°×360°大小的矩阵，存储到细胞数组中，插值方法结束。

4 构建数据的准确性分析

由于在插值方法训练时仅使用 80% 的实测数据，剩余的 20% 实测数据就可用于对方法进行准确性评估。这 20% 实测数据在训练完成后输入到神经网络中，比较实测值 $\boldsymbol{Y}_0$ 和神经网络计算出的预测值 $\boldsymbol{Y}$ 的差异来评估所构建数据的准确度。通常用标准误差（RMSE）和平均相对误差 (MRE) 来描述方法的精确度。

其中标准误差的计算公式为

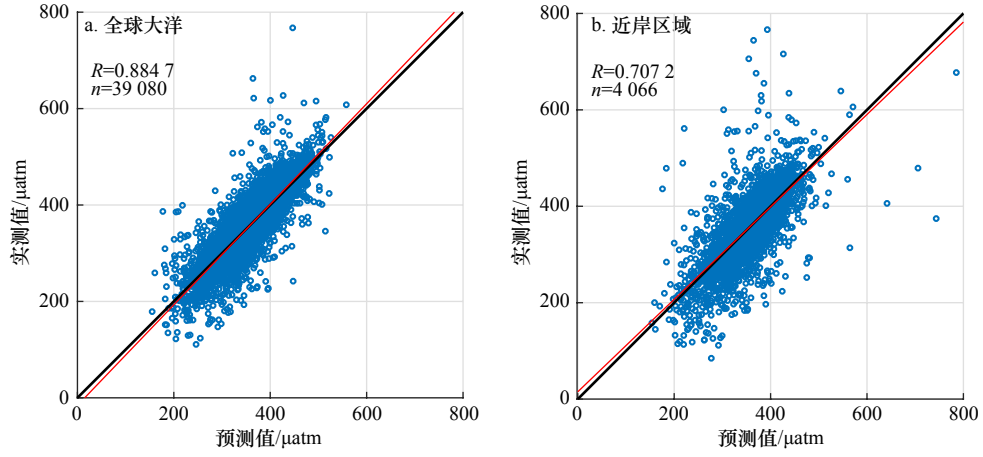
$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{Y}_{0i})^2}{n}}, \quad (14)$$

式中，$\boldsymbol{Y}_i$ 代表第 i 个样本的 $p\mathrm{CO}_2$ 神经网络预测值，$\boldsymbol{Y}_{0i}$ 代表第 i 个样本的 $p\mathrm{CO}_2$ 实测值，n 为进行误差评估的测试样本总数，参与神经网络训练的数据不用于误差评估。标准误差对一次预测中的特大或特小误差十分敏感，反映出预测值相对于实际值的偏离程度，是用于评价拟合效果的指标中最常用的。

平均相对误差的计算公式为对每个点误差占实测值的比重求平均：

$$\mathrm{MRE} = \frac{\sum_{i=1}^{n}\left|\frac{\boldsymbol{Y}_i - \boldsymbol{Y}_{0i}}{\boldsymbol{Y}_{0i}}\right|}{n}. \quad (15)$$

对比广义回归网络计算出的预测值 $\boldsymbol{Y}$ 和实测值 $\boldsymbol{Y}_0$ 可以发现，预测结果与实测数据基本一致。以 $\boldsymbol{Y}$ 为 x 轴，$\boldsymbol{Y}_0$ 为 y 轴作图，绝大部分数据点聚集在 $y=x$ 直线处，部分偏离较远但仍均匀地靠近直线并分布在直线两侧（图 5），回归线也十分逼近 $y=x$ 直线。全球大洋的预测值相较于实测值的平均相对误差为 2.97%，标准误差为 16.93 μatm，相关系数为 0.884 7。实测数据多的区域，如亚热带太平洋、赤道太平洋和亚热带大西洋，插值预测的表现最好，标准误差为 10.45～13.87 μatm，平均相对误差为 1.93%～2.66%。南太平洋数据量较少，误差却也很低，可能是因为数值变化范围较其他区域小，实测 $p\mathrm{CO}_2$ 值均在 250～450 μatm 之间，而不管哪个区域在这一区间内的预测值与实测值都很接近。表 2 给出了与其他方法的标准误差对比[11−13, 17−21]，在整体上，GRNN 略优于 FFNN 与 SOM，具体到特定区域范围时，一些机器学习算法的表现可能更加精确，例如 Chen 等[17] 使用随机森林算法重建了墨西哥湾的 $p\mathrm{CO}_2$ 变化，标准误差仅 9.10 μatm，然而仅适用于主导因素较为单一的小范围区域。也有研究将 SOM 和 FFNN 结合在一起，利用 SOM 将大西洋划分成多个区域，对每个区域训练一个 FFNN 来进行插值预测[20]，但精确度并没有显著提升，并且由于同时使用了两种神经网络，应用起来更加繁琐。近岸区域由于受到陆地径流和人类活动等因素影响，规律十分复杂，广义回归网络做出的预测表现与大洋区



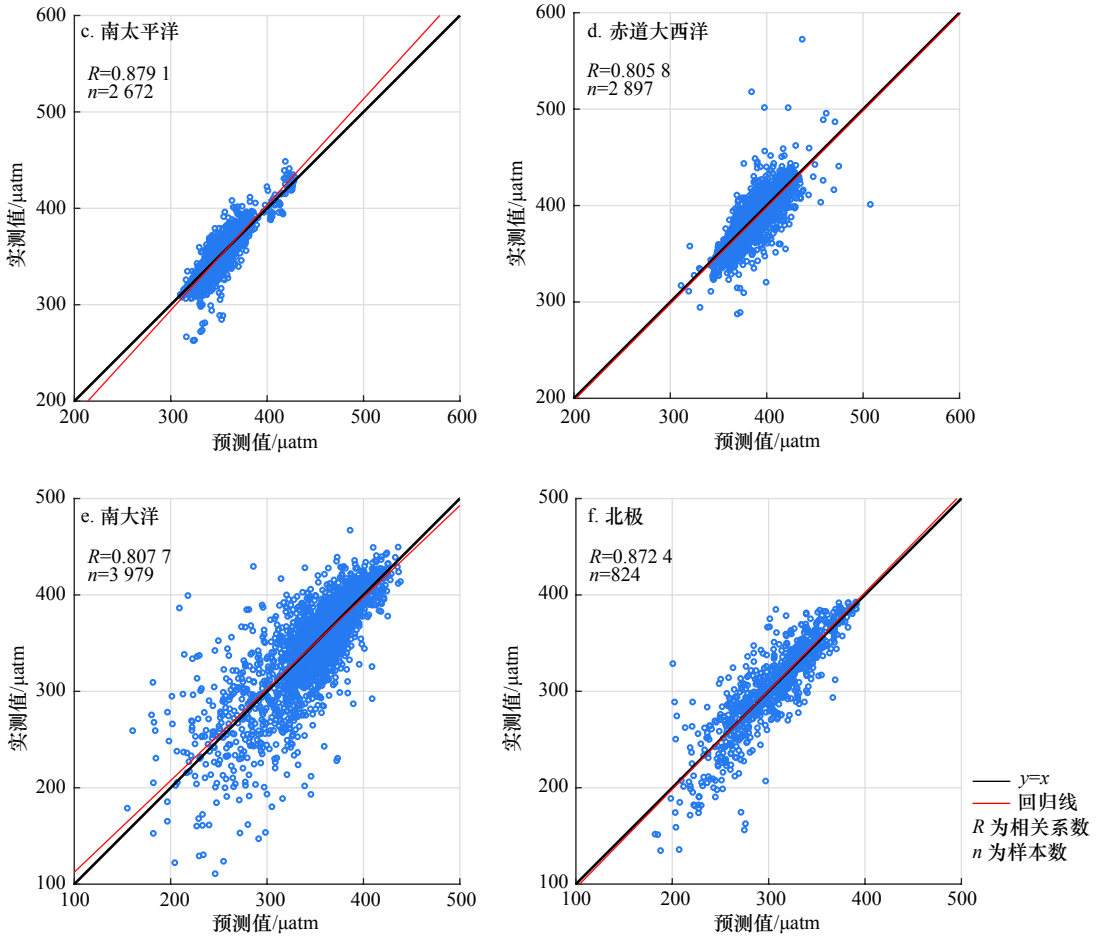


图 5　广义回归神经网络预测值与实测值对比

Fig. 5　Comparation of GRNN predict pCO_2 and *in situ* measurements

表 2　GRNN 与其他方法误差对比

Table 2　Comparation of errors between GRNN and other approaches

区域	标准误差/μatm				
	FFNN[13]	机器学习算法[17]	SOM[11, 12, 18-19]	SOM-FFNN[20-21]	GRNN
全球大洋	17.97	9.10(墨西哥湾)	17.60～20.20	20.00～22.80	16.93(2.97%)[b]
全球大洋及近岸[a]	–	–	–	–	21.60(3.52%)[b]
北极	22.05	–	–	–	23.92(6.00%)[b]
近极地大西洋	22.99	–	–	–	21.39(3.49%)[b]
近极地太平洋	34.77	–	–	–	24.57(4.59%)[b]
亚热带大西洋	17.28	–	–	–	13.87(2.66%)[b]
亚热带太平洋	15.86	–	–	–	11.50(2.15%)[b]
赤道大西洋	17.27	–	–	–	14.26(2.52%)[b]
赤道太平洋	15.73	–	–	–	10.45(1.93%)[b]
南大西洋	17.81	–	–	–	15.37(2.60%)[b]
南太平洋	13.52	–	–	–	9.97(2.04%)[b]
印度洋	17.25	–	–	–	11.64(2.25%)[b]
南大洋	17.40	–	–	–	24.59(4.87%)[b]
近岸区域[a]	–	–	–	42.40～48.00	46.87(8.83%)[b]

注: a表示此处近岸区域指水深小于200 m的海域; b表示括号内为本文方法的平均相对误差; –表示无数据; FFNN: 前反馈神经网络; SOM: 自组织映射神经网络; SOM-FFNN: 自组织映射神经网络与前反馈神经联用法; GRNN: 本文所使用的广义回归神经网络。

域相比较差，但相近于 Laruelle 等[21]使用 SOM-FFNN 法对近岸区域的预测表现。如果包含近岸区域，整体的标准误差将上升到 21.60 μatm，但其他的研究也只关注大洋区域，或者只关注近岸区域，并没有统一研究。

尽管存在一定的误差，GRNN 法的结果与 pCO_2 实测值的分布在高值和低值区域的位置上基本一致（图 6a，图 6b）。与同样使用 SOCAT 数据集的其他神经网络方法的数据产品进行对比（图 6b 至图 6d），图 6c 为 SOM 法，标准误差为 23.3 μatm，图 6d 为 SOM−FFNN 联用法，标准误差为 22.8 μatm，几种方法整体的季节

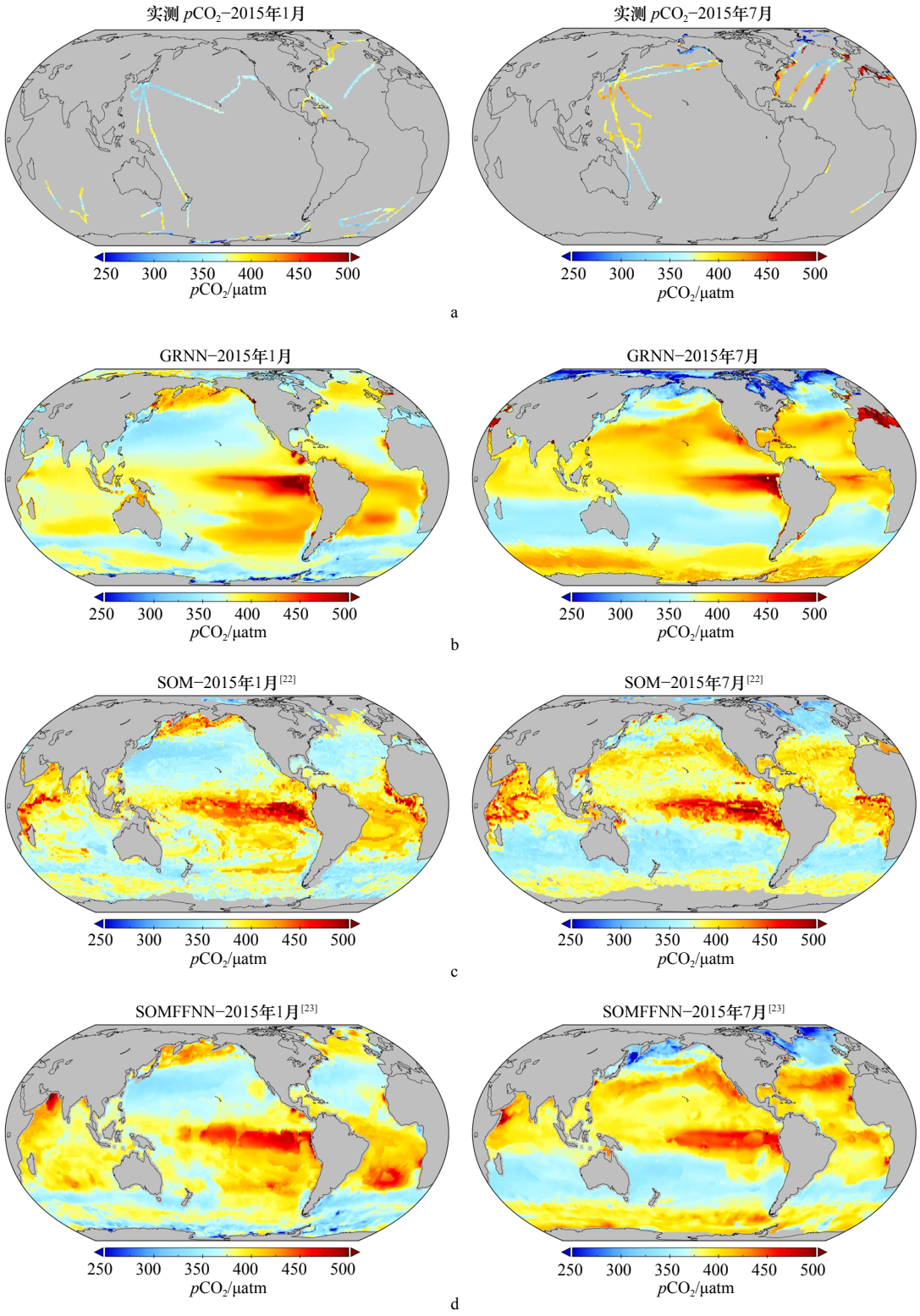


图 6 GRNN 与实测值及其他神经网络方法同时间点 pCO_2 数据结果对比

Fig. 6 Comparation of pCO_2 distribution between *in situ* measurements, GRNN and other approaches

趋势表现出高度的一致。在 1 月份, 南半球中纬度海域 pCO_2 高, 南太平洋东部高于西部; 北半球中纬度海域整体 pCO_2 低, 而北太平洋和北大西洋近极地区域与中纬度区域相反, pCO_2 高。7 月中纬度海域整体分布规律与 1 月大致相反, 南大洋 pCO_2 高, 北极区域低, 这些特征都与其他方法给出的结果相一致。尽管使用的实测 pCO_2 数据集一致, 不同方法使用的辅助参数种类也不同, 例如图 6 中另外两种方法中, 图 6c 使用的参数中没有经度, 使用了混合层深度, 图 6d 没有经纬度和时间, 这可能也是特定区域的 pCO_2 分布规律上各有差异的原因, 特别是印度洋等实测数据少的区域。此外, 不同研究使用的辅助参数空间覆盖范围不同, 导致构建出的 pCO_2 空间范围存在差异。不同类型的神经网络本身的特性也存在差异, 由于 SOM 给出的是离散的估计, 数据的空间连续性最差, 存在明显的斑块状分布。SOM-FFNN 虽然是连续的估计, 但是数据过渡也不太平滑, 锐利的边界仍存在。而表层海水并不是相互独立的, 由于物理混合过程的影响, 高分辨率的情况下相邻网格间月平均 pCO_2 不应该相差太大。比起其他方法, GRNN 法推演出的数据平滑程度更高, 不需再进行人为的二次处理来达到平滑过渡效果, 有潜力应用于更高分辨率的数据构建上, 如 0.25°×0.25°甚至更高。

除了神经网络法外, 与 Takahashi 等[24] 通过将数十年的实测数据校正到同一年, 构建气候态分布的方法对比, pCO_2 的整体分布规律也存在较高的一致性。如图 7b 是 Takahashi 等[24] 的研究中给出的校正到 2005 年的 1 月 pCO_2 全球分布, 图 7a 是同时间 GRNN 法给出的结果, 北太平洋的高值带和低值带、南大洋的低值区域等非常相似。尽管使用的源数据和方法本身上存在一些差异, 使不同研究的结果在具体的区域分布各有不同, 但整体的分布规律高度相似, 结合标准误差和平均相对误差来看, 有足够的理由相信广义回归神经网络在二氧化碳分压的插值预测上是可靠的。

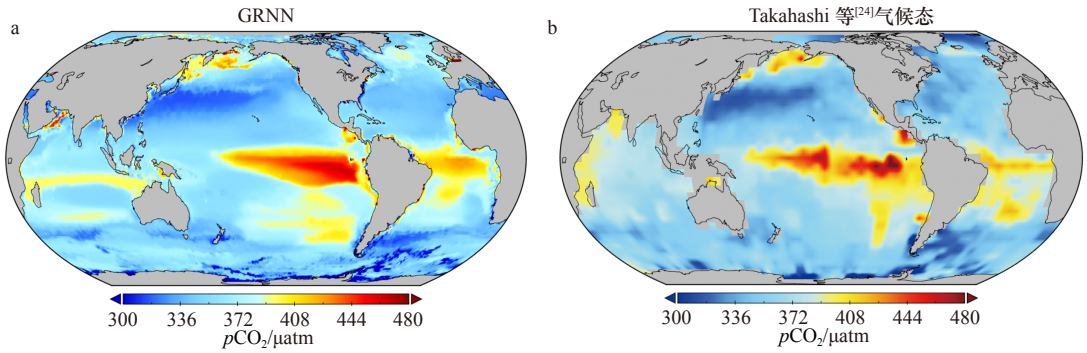


图 7　GRNN 法与 Takahashi 等[24] 气候态 pCO_2 数据对比(2005 年 1 月)

Fig. 7　Comparation of pCO_2 distributions between GRNN output and Takahashi[24] climatological mean (January, 2005)

5　结论

基于广义回归神经网络, 建立了表层海水二氧化碳分压与经纬度、时间、温度、盐度和叶绿素浓度间的非线性关系, 并据此重建了近 20 年来表层海水二氧化碳分压的全球分布, 标准误差为 16.93 μatm, 平均相对误差为 2.97%。与其他方法的对比证实了本插值方法的可靠性, 并且广义回归神经网络法的适用性更强, 对近岸区域也能做出预测, 其表现与只关注近岸区域的其他研究相近, 网络训练的速度也远高于其他神经网络。使用广义回归网络进行插值预测时, 由于不需要设定扩散速度外的参数, 插值结果的表现主要受实测数据本身影响。在参数选择方面, 输入参数对神经网络的预测表现影响很大, 但增加相关性低的参数并不能提高精确度, 反而会降低输出数据的平滑度, 导致分布呈现斑块状, 并且会增加计算时间。虽然增加相关性高的参数可以显著地提高精确度, 但受该参数可获得性的极大限制。例如本研究在 pCO_2 参数的构建中选用的叶绿素浓度, 该参数与 pCO_2 有较高的相关性, 但由于仅能获取该参数在 1998 年以后的大量卫星遥感数据, 也仅能用该参数重建 1998 年以后的 pCO_2 数据, 而由于无法获得足够的 1998 年之前的数据, 就无法用本研究建立的插值方法重建 1998 年之前的 pCO_2 数据。现有研究也大部分是通过使用卫星遥感数据来解决参数可获得性的问题, 但满足条件的卫星遥感数据也只有最近几十年的, 这也是大部分现有研究都只能重建近几十年 pCO_2 数据的原因。而更早期的观测数据过少, 很难支撑大范围的预测插值, 因此如何重建早期的 pCO_2 数据成为待解决的下一个难题。

参考文献：

[1] 曲宝晓, 宋金明, 袁华茂, 等. 东海海−气界面二氧化碳通量的季节变化与控制因素研究进展[J]. 地球科学进展, 2013, 28(7): 783−793.
Qu Baoxiao, Song Jinming, Yuan Huamao, et al. Advances of seasonal variations and controlling factors of the air-sea CO_2 flux in the East China Sea[J]. Advances in Earth Science, 2013, 28(7): 783−793.

[2] Takahashi T, Sutherland S C, Feely R A, et al. Decadal change of the surface water pCO_2 in the North Pacific: a synthesis of 35 years of observations[J]. Journal of Geophysical Research: Oceans, 2006, 111(C7): C07S05.

[3] Song Jinming. Biogeochemical Processes of Biogenic Elements in China Marginal Seas[M]. Berlin, Heidelberg: Springer Science & Business Media, 2010: 140−144.

[4] 宋金明, 李学刚, 袁华茂, 等. 渤黄东海生源要素的生物地球化学[M]. 北京: 科学出版社, 2019: 45−47.
Song Jinming, Li Xuegang, Yuan Huamao, et al. Biogeochemistry of Biogenic Elements in the Bohai Sea, Yellow Sea and East China Sea[M]. Beijing: Science Press, 2019: 45−47.

[5] Takahashi T, Sutherland S C, Wanninkhof R, et al. Climatological mean and decadal change in surface ocean pCO_2, and net sea-air CO_2 flux over the global oceans[J]. Deep-Sea Research Part II: Topical Studies in Oceanography, 2009, 56(8/10): 554−577.

[6] Takamura T R, Inoue H Y, Midorikawa T, et al. Seasonal and inter-annual variations in pCO_2^{sea} and air-sea CO_2 fluxes in mid-latitudes of the western and eastern North Pacific during 1999−2006: recent results utilizing voluntary observation ships[J]. Journal of the Meteorological Society of Japan, 2010, 88(6): 883−898.

[7] Sarma V V S S, Saino T, Sasaoka K, et al. Basin-scale pCO_2 distribution using satellite sea surface temperature, Chl *a*, and climatological salinity in the North Pacific in spring and summer[J]. Global Biogeochemical Cycles, 2006, 20(3): GB3005.

[8] Lefévre N, Watson A J, Watson A R. A comparison of multiple regression and neural network techniques for mapping in situ pCO_2 data[J]. Tellus B: Chemical and Physical Meteorology, 2005, 57(5): 375−384.

[9] Zeng J Y, Nojiri Y, Nakaoka S I, et al. Surface ocean CO_2 in 1990−2011 modelled using a feed-forward neural network[J]. Geoscience Data Journal, 2015, 2(1): 47−51.

[10] Zeng J, Nojiri Y, Landschützer P, et al. A global surface ocean fCO_2 climatology based on a feed-forward neural network[J]. Journal of Atmospheric and Oceanic Technology, 2014, 31(8): 1838−1849.

[11] Telszewski M, Chazottes A, Schuster U, et al. Estimating the monthly pCO_2 distribution in the North Atlantic using a self-organizing neural network[J]. Biogeosciences, 2009, 6(8): 1405−1421.

[12] Nakaoka S, Telszewski M, Nojiri Y, et al. Estimating temporal and spatial variation of ocean surface pCO_2 in the North Pacific using a self-organizing map neural network technique[J]. Biogeosciences, 2013, 10(9): 6093−6106.

[13] Denvil-Sommer A, Gehlen M, Vrac M, et al. LSCE-FFNN-v1: a two-step neural network model for the reconstruction of surface ocean pCO_2 over the global ocean[J]. Geoscientific Model Development, 2019, 12(5): 2091−2105.

[14] Körtzinger A. Determination of carbon dioxide partial pressure ($p(CO_2)$)[M]//Grasshoff K, Kremling K, Ehrhardt M. Methods of Seawater Analysis. 3rd ed. New York: Wiley, 1999: 149−158.

[15] Specht D F. A general regression neural network[J]. IEEE Transactions on Neural Networks, 1991, 2(6): 568−576.

[16] 陈明. MATLAB神经网络原理与实例精解[M]. 北京: 清华大学出版社, 2013: 208−237.
Chen Ming. MATLAB Neural Network Principle and Example Fine Solution[M]. Beijing: Tsinghua University Press, 2013: 208−237.

[17] Chen Shuangliang, Hu Chuanmin, Barnes B B, et al. A machine learning approach to estimate surface ocean pCO_2 from satellite measurements[J]. Remote Sensing of Environment, 2019, 228: 203−226.

[18] Friedrich T, Oschlies A. Neural network-based estimates of North Atlantic surface pCO_2 from satellite data: a methodological study[J]. Journal of Geophysical Research: Oceans, 2009, 114(C3): C03020.

[19] Hales B, Strutton P G, Saraceno M, et al. Satellite-based prediction of pCO_2 in coastal waters of the eastern North Pacific[J]. Progress in Oceanography, 2012, 103: 1−15.

[20] Landschützer P, Gruber N, Bakker D C E, et al. A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink[J]. Biogeosciences, 2013, 10(11): 7793−7815.

[21] Laruelle G G, Landschützer P, Gruber N, et al. Global high-resolution monthly pCO_2 climatology for the coastal ocean derived from neural network interpolation[J]. Biogeosciences, 2017, 14(19): 4545−4561.

[22] Zeng Jiye, Matsunaga T, Saigusa N, et al. Technical note: Evaluation of three machine learning models for surface ocean CO_2 mapping[J]. Ocean Science, 2017, 13(2): 303−313.

[23] Landschützer P, Gruber N, Bakker D C E. Decadal variations and trends of the global ocean carbon sink[J]. Global Biogeochemical Cycles, 2016, 30(10): 1396−1417.

[24] Takahashi T, Sutherland S C, Chipman D W, et al. Climatological distributions of pH, pCO_2, total CO_2, alkalinity, and $CaCO_3$ saturation in the global surface ocean, and temporal changes at selected locations[J]. Marine Chemistry, 2014, 164: 95−125.

A general regression neural network approach to reconstruct global 1°×1° resolution sea surface pCO_2

Zhong Guorong [1,2,3,4], Li Xuegang [1,2,3,4], Qu Baoxiao [1,3,4], Wang Yanjun [4], Yuan Huamao [1,2,3,4], Song Jinming [1,2,3]

(1. *Key Laboratory of Marine Ecology and Environmental Sciences*, *Institute of Oceanology*, *Chinese Academy of Sciences*, *Qingdao* 266071, *China*; 2. *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*; 3. *Marine Ecology and Environmental Science Laboratory*, *Pilot National Laboratory for Marine Science and Technology (Qingdao)*, *Qingdao* 266237, *China*; 4. *Center for Ocean Mega-Science*, *Chinese Academy of Sciences*, *Qingdao* 266071, *China*)

Abstract: Sea surface partial pressure of carbon dioxide (pCO_2) is a crucial parameter for estimating ocean carbon source and sink term, but its sparse and uneven in situ measurements in space and time lead to large uncertainty in the estimate of sea-air CO_2 flux and characteristics of ocean carbon source and sink. To eliminate this uncertainty, a general regression neural network approach using the Surface Ocean CO_2 Atlas (SOCAT) dataset, based on the non-liner regression of pCO_2 and longitude, latitude, time, temperature, salinity and concentration of chlorophyll, was successfully used in the reconstruction of global 1°×1° resolution monthly sea surface pCO_2 from 1998 to 2018, with a root mean square error (RMSE) of 16.93 μatm and a mean relative error (MRE) of 2.97%, lower than existing feed-forward neural network (FFNN), self-organizing neural network (SOM) and machine learning approaches. The global distribution of pCO_2 obtained by this approach agrees well with existing researches.

Key words: general regression neural network; sea surface pCO_2; global ocean grid data