J. Lake Sci.(湖泊科学), 2020, **32**(3): 877-886
DOI 10.18307/2020.0326

基于同步观测信息利用的高寒湖泊湍流通量数据插补方法*

靳　铮[1,2],张雪芹[1**]
(1:中国科学院地理科学与资源究所,中国科学院陆地表层格局与模拟重点实验室,北京 100101)
(2:中国科学院大学,北京 100049)

摘　要: 源区划分和质量过滤提高湖面涡动相关通量数据可靠性的同时,却降低了通量时间序列的连续性. 为此,本文基于TensorFlow机器学习框架构建了一种超宽人工神经网络(ANN)模型. 在选择输入ANN模型的特征变量信息时,我们采取了尽可能获取湍流输送过程中热力、动力学同步观测背景强迫信息的原则. 通过ANN模型模拟通量的插补,本文实现了通量时间序列连续性的优化,插补后的羊卓雍错湖面通量数据的时间覆盖率从不足0.40提升至超过0.98. 基于10次折叠交叉验证的ANN模型通量模拟性能检验则表明,各个检验组之间ANN模型的模拟误差波动较小,这显示出了较好的稳健性. 具体地讲,感热通量、潜热通量和水汽通量原始观测平均值分别约为18.8 W/m^2、81.5 W/m^{2}和1.84 mmol/(s·m^2),10组交叉验证的插补感热通量、潜热通量和水汽通量平均绝对误差分别为5.4 W/m^2、15.7 W/m^{2}和0.35 mmol/(s·m^2). 这表明本文所探索的ANN建模结构和同步观测变量筛选原则可更充分地利用观测点局地同步观测信息估算通量强度,有效地优化湍流通量数据的时间连续性,从而提升通量数据的可分析性.
关键词: 数据插补;人工神经网络;通量观测;涡动相关方法;羊卓雍错

Optimization method for alpine lake turbulent flux data based on micro-meteorological information utilization*

JIN Zheng[1,2] & ZHANG Xueqin[1**]
(1: *Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing* 100101, *P.R.China*)
(2: *University of Chinese Academy of Sciences, Beijing* 100049, *P.R.China*)

Abstract: Source area partition and quality filtering can improve the dependability of eddy covariance (EC) flux data while reducing its temporal consistency. Here, we constructed an ultra-wide artificial neural network (ANN) structure based on the TensorFlow framework. For the ANN inputting feature information selection, we attempted to establish feature vectors utilizing adequate thermodynamic forcing information of micro-meteorological background. The temporal consistency of EC data was optimized by interpolating with the ANN modeled fluxes, raising the temporal coverage rates from under 0.40 to over 0.98 for the flux data at the lake surface of Yamzhog Yumco. The evaluation of flux simulation performance via 10-fold cross-validation indicates that the bias level exhibits minuscule perturbation over different subsamples, disclosing preferable robustness for the ANNs model. Comparing for the approximately 18.8 W/m^2/81.5 W/m^2 of average value for the observed sensible/latent heat flux, 1.84 mmol/(s·m^2) for water vapor flux, the mean absolute errors is 5.4 W/m^2 for the simulated sensible heat flux, 15.7 W/m^2 for the simulated latent heat flux, and 0.35 mmol/(s·m^2) for the water vapor flux. The results suggest that the combination of ANN structure with variable selecting principle can utilize the micro-meteorological information of field observation more sufficiently to estimate the flux intensity. Consequently, the temporal consistency is efficiently optimized with the analysability of EC flux data enhanced. The optimization method we proposed makes the interpolation of EC flux observation data no longer depend on the calculation of specific micro-meteorological elements such as turbulence transport coefficient. The paper provides a reference idea for improving the data quality of EC flux observation experiments for alpine lakes and other harsh environments.

---

* 2019-05-27 收稿;2019-12-03 收修改稿.
国家自然科学基金项目(41471064)资助.
** 通信作者;E-mail:zhangxq@igsnrr.ac.cn.

Keywords: Data interpolation; artificial neural network; flux measurement; eddy covariance method; Yamzhog Yumco

涡动相关(eddy covariance, EC)方法在国内外湖泊水热和成分通量观测研究中均有广泛应用[1-6]. 太湖在2013年就建成了由6个通量站组成的中尺度涡动相关通量观测网[3],并利用观测数据分析和模拟了该地区湖气间能量、水热相互作用. 2014年鄱阳湖的涡动相关研究[5]为大型浅水湖湖面能量通量划分了不同的水文过渡期格局. 洱海[7]和纳木错[8]的涡动相关观测研究初步揭示了高寒湖泊的水面湍流通量特征. 受制于仪器性能和观测下垫面的异质性,EC观测的通量数据时间序列会产生缺测点[9]. 缺测点产生的原因主要包含信号异常、仪器损坏以及湍流平稳性统计未能满足通量计算假设等[10]. 特别是高寒湖泊EC观测面临着封冻期维护、强风和低温等问题. 因此,EC系统一般采取岸基或近岸安装[11-13]. 这使得EC系统所测来自陆面源区的通量数据和湖面源区的混杂一起. 通量源区划分并提取目标源区通量数据后,EC观测的有效数据比例显著下降. 湖面源区通量数据有效数据时间分布存在显著的不均匀现象,从而严重影响EC通量数据的可分析性,因此有必要对EC通量数据进行插补. 湖面水热通量观测研究中,学者们把温度梯度、风速、水汽压差等作为湍流传输过程的基本物理背景信息来估算通量强度(如Bulk Aerodynamic Transfer Model[13]和DYRESM[14]). 这类种估算方式除了考虑动力、热力机制之外,还需考虑物理量纲的对应,以便将估算模型以等式的形式表达出来. 但是湍流输送过程的时空尺度决定了这类梯度量纲模型不足以描述湍流传输在其宏观特征背后的微观过程. 除此之外,通量数据的插补还可以使用查表法[15]、动态线性回归方法[16]和非线性回归方法[17]等基于统计原理的方法. 这类方案是根据观测经验,以温度梯度、水汽压等关键背景物理信息为依据,通过拟合计算得到对通量强度的最优统计估计来完成插补的. 相对于上述方案而言,ANN(artificial neural network,人工神经网络)方法可以更加充分地利用羊卓雍错观测实验中较丰富的辅助观测系统和较长观测时间带来的数据量优势. 由于羊卓雍错的特殊近地层湍流背景[18],湍流输送过程在此处与其背景动力、热力强迫过程的关系更加复杂,非线性特征更显著. 故本文在温度梯度、风速、水汽压差等核心要素的基础上,将更多的环境气象要素同时引入到湍流传输物理背景信息的描述中,并使用ANN模拟当前信息组合与通量强度值的映射关系.

尽管以ANN为基本工具的机器学习方法广泛应用于地球科学研究中[19],但其在湖泊通量观测的数据插补方面仍有待加强. 洱海EC通量观测实验[7]使用了一个神经元总数20个以内的ANN模型来模拟和插补通量数据,其ANN模型输入变量包含风速、水面温度、气温、饱和水汽压差4项. 但该研究对ANN模型估算通量强度的准确性未作进一步讨论. 由于ANN模型的性能取决于数据所包含的有效信息量和噪声水平[20],故其模拟性能的检验是必要的. 我们从信息利用的角度设计了一种针对EC湍流通量数据的ANN模拟插补及验证方案. 通量模拟输入信息中增加变量时,信噪比较低的变量会影响ANN模型的性能,同时大幅增加ANN训练时间. 为了给输入ANN模型的大量有效信息和噪声组合提供充足的概率样本空间定义域,本文采用了超宽ANN结构[21]. 基于超宽结构的ANN模型神经元总数超过2000个,本研究利用并行计算技术解决了模型训练的时效问题. 本文对上述超宽结构ANN模型的通量估算性能进行了交叉验证和误差分析,提供了客观且全面的验证结果和误差分析结论,并讨论了EC湍流通量数据插补问题中ANN模型的泛化,为高寒湖泊湍流通量观测研究的数据插补提供了一种参考思路.

1 数据与方法

1.1 羊卓雍错涡动相关观测

羊卓雍错湖湖面面积约为643 km^2[22],是喜马拉雅山脉北麓最大的内陆封闭湖泊. EC设备及辅助观测系统的设置地点为羊卓雍错湖靠近浪卡子县白地水文站的近岸浅滩(29°07′28″N, 90°26′27″E),海拔高度4420.63 m. 为避免被湖泊封冻期前后的冰凌冲击摧毁,通量设备于每年的1月上旬进行拆卸,3月下旬重新安装. 本研究采用通量系统2016年和2017年非封冻期同一观测时段(4月3日0:00—12月31日23:30)的水热通量、湍流、气温、湿度、水面温度和水面辐射等同步观测数据. 此外,2016年3月30日—4月2日以及2018年1月1—4日的数据将用于测试ANN模型的时间扩展性能,不参与ANN模型的训练及交叉验证. EC设备信息与数据采集的详细情况见参考文献[20].

1.2 神经网络建模与检验方法

1.2.1 同步观测变量的特征工程　特征工程是对 ANN 输入数据进行标准化处理的过程[23],原始同步观测数据经过特征工程转换为可直接输入 ANN 的特征向量. 特征向量的处理方式对 ANN 的训练时效和模拟性能均有重要影响. 本文使用 ANN 方法进行通量强度与同步观测特征间的映射关系拟合,输入 ANN 的同步观测特征变量有 12 个:气温、水面温度、平均水平风速、平均湍流动能、莫宁-奥布霍夫稳定度参数[24]、气压、饱和水汽压、水汽压和四分量长、短波辐射. 上述加入输入数据样本的特征变量除了与涡动通量强度观测值同一时刻的 30 min 平均值外,还包含它们前后各 30 min 的平均值,故 ANN 的输入特征向量维度为 36. 同步观测特征变量的标准化处理采用平移缩放法,即根据各个变量的强度概率密度分布剔除异常值后平移并缩放至[0, 1]区间. 训练样本由处理完成后的同步观测变量特征向量及观测通量强度序列组成. 由于 ANN 模型具备较强的非线性映射拟合性能[25],本研究特征变量选取时并不考虑所选变量和通量强度的统计相关性,而只需要特征变量与湍流输送过程的热力、动力过程存在理论关联即可.

1.2.2 建模框架与参数调试　本文使用 Google®的开源机器学习框架 TensorFlow[26]进行 ANN 模型参数化构建及训练,并使用 Keras[27]开源 Python 模块对 TensorFlow 的功能框架进行调用. 为了保证映射关系拟合计算的时效性,本文采用 CUDA®(Nvidia®, Inc.)并行计算方案,硬件计算单元型号为 GTX-1080®,峰值性能为每秒 9 万亿次单精度浮点运算(9 TFLOPS). ANN 的初始化模型采用密集连接结构,不加入偏差项神经元,损失函数为训练样本的平均绝对误差,拟合方式为随机梯度下降算法[28],并利用随机参数剔除[29]方案弱化过拟合效应. ANN 的激活函数采用 PReLU[30](Parametric Rectified Linear Unit)和 Tanh[31](Hyperbolic Tangent),Tanh 函数用于激活 ANN 的第一个隐藏层,PReLU 函数则用于激活 ANN 的输入层和连接输出层的隐藏层. 已有研究表明隐藏层宽度于一定的范围内增加时,ANN 可以对非线性过程作出近似与牛顿迭代法的线性求解[21]. 所以,本文参数调试采用隐藏层宽度指数递增搜索的策略,从初始化的 16 个隐藏层神经元开始,逐步增加隐藏层神经元至 2^n 个,确定了合适的隐藏层数与神经元数后,再对学习率、剔除率和训练次数等参数进行调优.

1000 h(计算单元满载运行时间)左右的参数调试结果表明:浅层超宽密集连接结构的 ANN 对涡动通量具有较好的模拟效果. 表 1 为 Keras 机器学习模块下 ANN 设置参数的搜索结果. ANN 的两个隐藏层神经元数量分别达到 2048 和 1024 个,这种超宽结构可以为不同质量、不同量纲的输入物理变量组合提供较大的概率空间,从而在迭代过程中搜索更多的映射关系,实现输入信息的充分利用. 由于感热通量、潜热通量和水汽通量所映射的同步观测特征是相同的,而通量强度数据的噪声不同. 故上述 3 种通量的拟合采用相同的 ANN 结构,只是利用不同的学习率和迭代次数来适应它们的噪声水平. 搜索计算所得的 ANN 模型虽然只有两个隐藏层,但两个隐藏层宽度均达到 2^{12} 个神经元后 ANN 的权重矩阵包含的权重数量已经超过 1.6×10^7 个,一次模型拟合所需训练时间超过 12 h. 3109 个神经元的 ANN 模型在对应的训练迭代次数(表 1)上已经接近收敛,每 100 次迭代的梯度波动率小于 0.01%. 考虑到实验的时效性及模型的收敛情况,我们未继续进行更大规模的 ANN 拟合计算.

表 1 ANN 参数设置的搜索结果

Tab.1 Searching results of parameters set for the proposed ANN

参数类型	参数值		
	感热通量	潜热通量	水汽通量
输入层神经元数		36	
隐藏层 1 神经元数		2048	
隐藏层 2 神经元数		1024	
输出层神经元数		1	
输入层激活函数		PReLU	
隐藏层 1 激活函数		Tanh	
隐藏层 2 激活函数		PReLU	
输出层激活函数		Linear	
隐藏层 1 剔除率		0.5	
平均学习率	0.0436	0.0547	0.0542
学习率衰减率		10^{-6}	
平均迭代次数	63650	91975	90725

* 平均学习率和平均迭代次数为 10 批次扰动训练的平均值.

1.2.3 折叠交叉验证　本文采用折叠交叉验证方法[32]对机器学习算法的性能进行了评估,并选取了 10 次折叠的交叉验证方案. 首先,我们按通量数据的强度概率密度分布将训练样本分割为 10 个折叠子样本,使各折叠子样本的通量数据均具有和原始样本相近的强度概率密度分布和平均值,并且在观测时段内均匀排列.

ANN 模型拟合计算时使用其中 9 个子样本合并进行训练,保留 1 个子样本用于验证模拟效果,然后交换用于验证的子样本. 在对 10 个子样本分别进行效果检验后,统计其平均结果和波动情况,以得到模型的误差期望和稳定性估计.

在搜索计算试验中,虽然 10 组训练子样本数据所驱动的 ANN 结构及拟合参数是相同的,但是由于 ANN 拟合计算采用了随机梯度下降方法,拟合后所得的模型权重参数在每一组交叉验证模型中存在微小差异,所以有必要分析各交叉样本模型性能的稳定性. 基于湍流输送过程的高度非线性[33]和 ANN 模拟输出可能产生的正负误差对称性[34],本文基于 10 组经过验证的 ANN 模型,额外增加了 10 批次的拟合计算试验(共 100 个 ANN 拟合扰动模型). 我们对每一批次 ANN 模型的拟合参数进行了微调扰动,训练了共计 110 个 ANN 模型作为备选集合成员. 而后以平均绝对误差的方差最低为标准,我们选取了 10 个成员对通量强度进行集合模拟取平均,以进一步降低最终模拟结果的不确定性.

2 羊卓雍错通量数据插补及验证

2.1 湖面水热通量数据插补

表 2 插补前后湖面通量的时间覆盖率

Tab.2 Temporal coverage rate of lake surface fluxes before/after the interpolation

时期与状态		时间覆盖率/%		
		感热通量	潜热通量	水汽通量
2016 年	插补前	38.3	26.4	26.4
	插补后	98.5	98.5	98.5
2017 年	插补前	36.7	36.5	36.5
	插补后	99.2	99.2	99.2

插补前后通量强度数据的时间覆盖率差异显著(表 2). 湖面通量与陆面通量源区划分以及低质量等级数据剔除后,观测期间通量数据时间覆盖率均降至 40%以下. 由于仪器故障,2016 年的潜热通量和水汽通量数据从 10 月中下旬开始大量出现异常状况(如水汽瞬时数浓度测定值为负). 为减缓仪器故障对数据可靠性和整体噪声造成的影响,本文剔除了 2016 年 9 月 30 日以后的全部潜热通量和水汽通量数据. 故 2016 年的潜热通量与水汽通量时间覆盖率(26.4%)相对感热通量而言更低,同时也低于它们 2017 年同期的水平. 通量数据插补后仍有少量缺测(比例低于 2%),这与输入 ANN 模型的同步观测特征值缺测或异常有关.

2.2 折叠交叉验证

10 组折叠交叉验证的结果(表 3)显示:ANN 模型的通量模拟性能十分稳定,各个交叉验证分组之间的误差波动较小. 对称性平均绝对百分比误差(*SMAPE*)[35]作为验证模型性能的关键指标,感热通量 10 组交叉验证中的平均值为 31.871%,潜热通量与水汽通量的交叉验证中的平均值分别为 20.8%和 20.7%,这表明 ANN 模型对潜热通量和水汽通量的模拟效果优于感热通量. 这种差别产生的原因是感热通量观测数据的噪声水平高于潜热通量和水汽通量. 平均绝对误差(*MAE*)在 3 种通量的各 10 个交叉验证组中的平均值为 5.4 W/m^2、15.7 W/m^2和 0.35 $mmol/(s\cdot m^2)$. *MAE* 与通量观测平均值的百分比可作为通量模拟的期望误差. 在感热通量、潜热通量和水汽通量的观测平均值分别为 18.8 W/m^2、81.5 W/m^2和 1.84 $mmol/(s\cdot m^2)$的情况下分别为 28.7%、19.3%和 19.2%. 半小时分辨率下,潜热通量和水汽通量的模拟期望误差相对较小. 对于通量强度的观测期平均值而言,感热通量、潜热通量和水汽通量的期望误差分别为 2.0%、1.3%和 1.8%. 这由它们在 10 个验证组中各组的观测值平均与模拟值平均计算得出. 通量模拟的整体平均值误差期望远小于半小时分辨率单个数值的误差期望,原因是通量模拟误差的正负对称性较好(图 1),参与平均计算的模拟值越多,整体平均值模拟误差就于 0 的附近越稳定.

10 次交叉折叠验证表明,*MAE* 差别最大的验证组之间回归分布状况却十分接近,模型性能在 10 个交叉验证组间的波动很小(图 2). 这反映了 10 组折叠交叉验证的样本分割具有良好的均一性. 模拟误差在不同的通量强度下的分布不均匀,感热通量对于 0~15 W/m^2之间时模拟误差明显更小,潜热通量和水汽通量于数值较大时模拟误差将增大. 这种现象出现的原因是超宽 ANN 对数据量有较高的敏感性,通量强度数据的概率分布在一定程度上符合正态分布,高值和低值数据所占比例相对于平均值附近的数据更少. 由对称绝对百分比误差(*SAPE*)的分位数分布(图 3)可见,*SAPE* 的平均值于各个通量的 10 组交叉验证中均大于其中

位数,表明多数情况下模拟误差小于模型误差期望.

表 3 ANN 模型的 10 组折叠交叉验证结果*

Tab.3 Results of 10-fold cross-validation on ANN models

统计参数	感热通量									
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
R	0.79	0.81	0.80	0.83	0.85	0.80	0.84	0.83	0.82	0.80
SMAPE/%	32.72	32.07	31.21	31.34	31.17	33.56	31.11	30.62	31.46	33.45
MAE/(W/m^2)	5.43	5.43	5.55	5.15	5.50	5.61	5.15	5.22	5.34	5.75
AVG-Obv/(W/m^2)	18.72	18.69	19.00	18.85	19.02	18.67	18.75	18.77	19.04	19.00
AVG-Est/(W/m^2)	18.31	19.00	18.80	18.54	19.65	18.30	18.56	18.78	18.53	17.99
	潜热通量									
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
R	0.81	0.76	0.78	0.76	0.81	0.79	0.81	0.77	0.77	0.77
SMAPE/%	20.65	20.49	20.68	20.96	20.48	20.55	20.92	21.18	21.03	21.74
MAE/(W/m^2)	15.23	15.69	15.55	15.76	15.83	15.48	15.75	15.83	16.13	16.38
AVG-Obv/(W/m^2)	80.44	80.74	81.36	81.29	82.40	81.68	81.66	81.82	82.18	81.47
AVG-Est/(W/m^2)	81.02	79.25	81.19	78.24	82.13	79.82	82.26	82.68	81.15	82.73
	水汽通量									
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
R	0.77	0.77	0.76	0.76	0.79	0.77	0.82	0.80	0.77	0.76
SMAPE/%	20.24	21.08	21.80	20.70	19.92	19.62	20.39	20.77	21.61	21.47
MAE/(mmol/(s·m^2))	0.34	0.35	0.37	0.36	0.34	0.32	0.34	0.35	0.37	0.37
AVG-Obv/(mmol/(s·m^2))	1.84	1.82	1.82	1.86	1.85	1.83	1.84	1.84	1.85	1.84
AVG-Est/(mmol/(s·m^2))	1.82	1.80	1.77	1.80	1.80	1.84	1.90	1.82	1.82	1.82

* *R* 为 Pearson 相关系数[36], *SMAPE* 为对称性平均误差百分比, *MAE* 为平均绝对误差, *AVG*-Obv 为观测值平均值, *AVG*-Est 为模拟值平均值, F1~F10 为 10 个折叠组的编号.

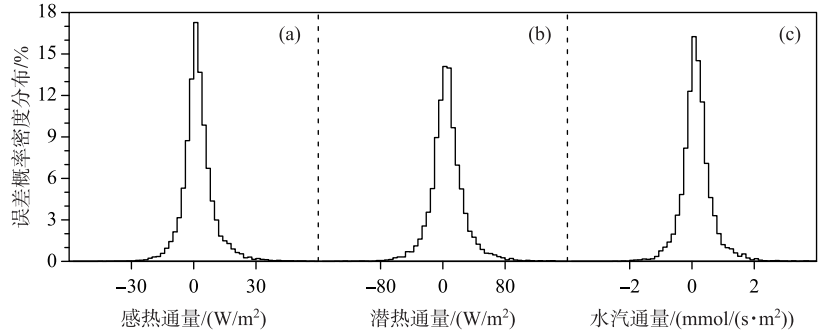

图 1 ANN 模型模拟误差的概率密度分布

Fig.1 Probability distributions of bias in ANN estimated fluxes

本文在计划修复的观测期(4 月 3 日 0:00－12 月 31 日 23:30)之外截取了两段未参与 ANN 模型拟合的数据,对比了 ANN 模型的模拟结果和观测值(图 4),结果显示 ANN 模型对潜热通量的模拟于 30~120 W/m^{2}的范围内模拟效果极好,超出该范围则误差显著增加,但变化趋势的一致性仍然较高,这与上文中 ANN 模型的折叠交叉验证结果相符.

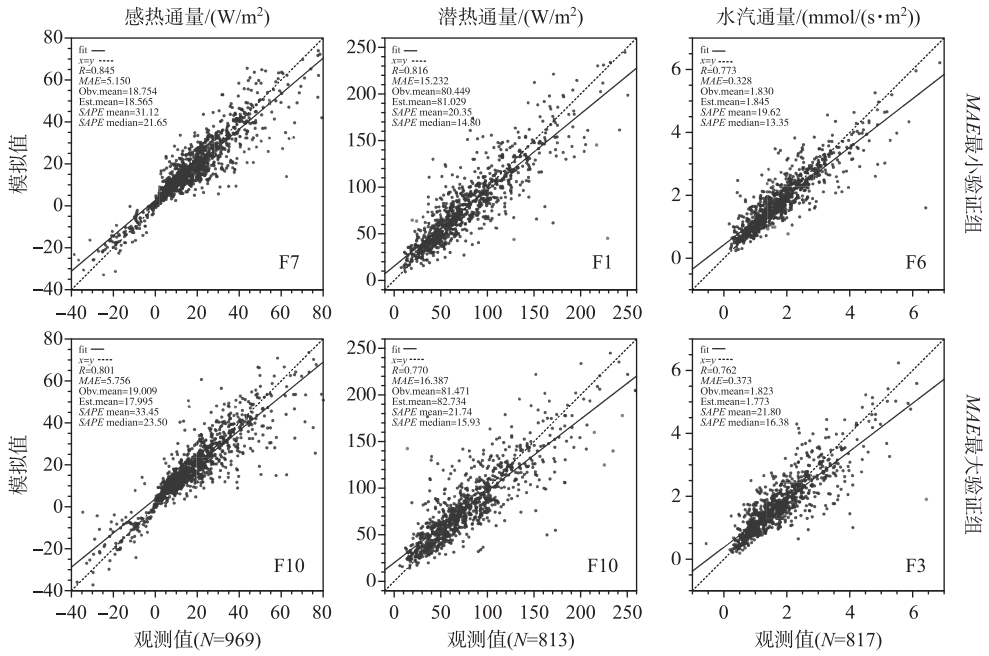

图 2 10 次折叠交叉验证中平均绝对误差最大和最小的验证组的回归分析结果对比
(回归结果均通过了 99.9% 的置信区间检验,F1、F3、F6、F7 和 F10 为折叠组的编号)

Fig.2 Comparisons between the minimum and maximum mean absolute error of regression analysis results in 10-fold cross-validation groups

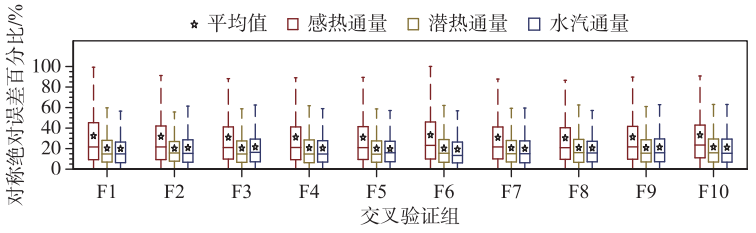

图 3 10 个交叉验证组的对称性绝对百分比误差分位数箱线图(F1~F10 为 10 个折叠组的编号)

Fig.3 Quantile boxes of symmetric absolute percentage error in 10-fold cross-validation groups

2.3 水热通量特征分析

2016 年与 2017 年湖面通量数据的缺失状况较为一致(图 5). 整个观测期内每天 0:00－9:00 是湖面通量数据时间覆盖率最低的时段. 通量源区定位取决于风速和风向,这表明该时间段内的主风向为陆地指向湖面. 与此相反,9:00－18:00 时段湖面通量的时间覆盖率则远大于观测期平均水平. 上述现象表明观测地点存在显著的湖陆风循环.

经 ANN 模型插补,湖面通量于观测期间的变化特征得以清晰而完整地呈现(图 5). 感热通量 2016 年与 2017 年的状况总体相似. 4－6 月有着稳定而显著的日变化,每天的峰值多于 10:00－16:00 间出现,其中 5 月全段和 6 月上半月的日变化幅度较大. 从 7 月下旬起日变化减弱,但每天的整体强度显著增大. 10 月下旬—11 月底出现了持续一个月的高峰,且高峰期整体强度上 2017 年大于 2016 年. 高峰期过后迅速减小至 7 月前的水平.

潜热通量最为显著的特征是日变化,每天的峰值于 12:00－18:00 间出现. 2016 年和 2017 年的 10－12 月有着近两个月的高峰期. 与感热通量相反,2016 年的潜热通量高峰期整体强度大于 2017 年. 由于观测期

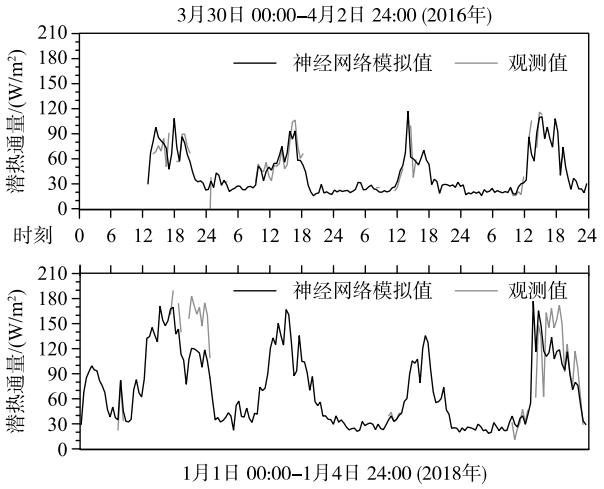

图 4 ANN 模拟的潜热通量与观测值的对比

Fig.4 Comparison between ANN estimated and observed latent heat flux

间气温变化幅度小于 20℃,此幅度下温度变化对汽化潜热的影响低于 3%,故水汽通量和潜热通量的强度变化几乎一致. 从两年的观测其整体状况来看,湖泊的能量释放于 2016 年 4 月初－10 月末是一个渐强的过程,10 月末达到顶峰. 而 2017 年的能力释放顶峰延后至 11 月下旬.

根据算法插补后的水汽通量数据及 ANN 的模拟误差期望,羊卓雍错湖面蒸发量的 EC 观测结果于 4－12 月为 740±9 mm(2016 年)和 703±7 mm(2017 年). 观测期湖面的感热通量平均值为 14.9±0.2 W/m^2(2016 年)和 14.3±0.2 W/m^2(2017 年),潜热通量平均值为 78.9±1.5 W/m^2(2016 年)和 74.4±1.5 W/m^2(2017 年),水面净辐射的平均值为 126.7 W/m^2(2016 年)和 139.5 W/m^2(2017 年). 由此可知湖面在 2016 年和 2017 年观测期间的能量平衡状况均为净流入,净流入能量为 7.737×10^8 J/m^2(2016 年)和 1.198×10^9 J/m^2(2017 年). 在不考虑湖水热力层结动态机制的情况下,净流入的热量可使湖面下方 20 m 深的湖水平均每月上升约 1~1.5℃,在水体的垂直动态热平衡下,不同深度的水温变化可能各不相同.

3 结论与讨论

针对高寒湖泊涡动通量观测有效数据比例偏低的问题,本研究利用了近年来机器学习研究中 GPU 并行计算和算法优化方面的发展成果,通过基于信息利用原则的超宽 ANN 模型,有效优化了 EC 通量数据的连续性,并利用 10 次折叠交叉验证方法检验了 ANN 模拟的效果. 主要结论如下:

在 ANN 模型插补通量数据中有效地利用了同步观测特征信息与通量强度的热力、动力关联性. 从数据插补效果看,该方案揭示了湍流通量回归中跨量纲、跨维度映射拟合的可行性和有效性. 羊卓雍错湖面感热通量、潜热通量和水汽通量的有效观测平均值分别为 18.8 W/m^2、81.5 W/m^2 和 1.84 mmol/(s·m^2). 根据 TensorFlow 机器学习框架下超宽结构 ANN 模型交叉验证结果,模拟的 *AME* 分别仅有 5.4 W/m^2、15.7 W/m^2 和 0.35 mmol/(s·m^2). 10 个交叉验证分组之间的误差波动幅度分别不超过 1 W/m^2、2 W/m^2 和 0.05 mmol/(s·m^2). 交叉验证组中全体观测值平均与模拟值平均的期望误差分别为 2.0%、1.3% 和 1.8%. 这表明超宽 ANN 模型的通量模拟性能十分稳定,在各个交叉验证组中波动较小,且模拟误差具有较好的正负对称性.

高寒湖泊环境带来了湍流通量观测的缺测问题,通量数据插值是观测研究的必要部分. 湍流通量过程具有高复杂度、高非线性的特征,观测数据量大、种类多. ANN 模型对输入数据的原始形态没有任何限制,无论所测同步观测数据的时间同步性、空间尺度、量纲、精度、采样率、噪声水平存在怎样的差别,都可以通过特征工程的标准化处理输入 ANN. 超宽结构和 GPU 并行计算同时保证了 ANN 模型模拟通量时输入数据的

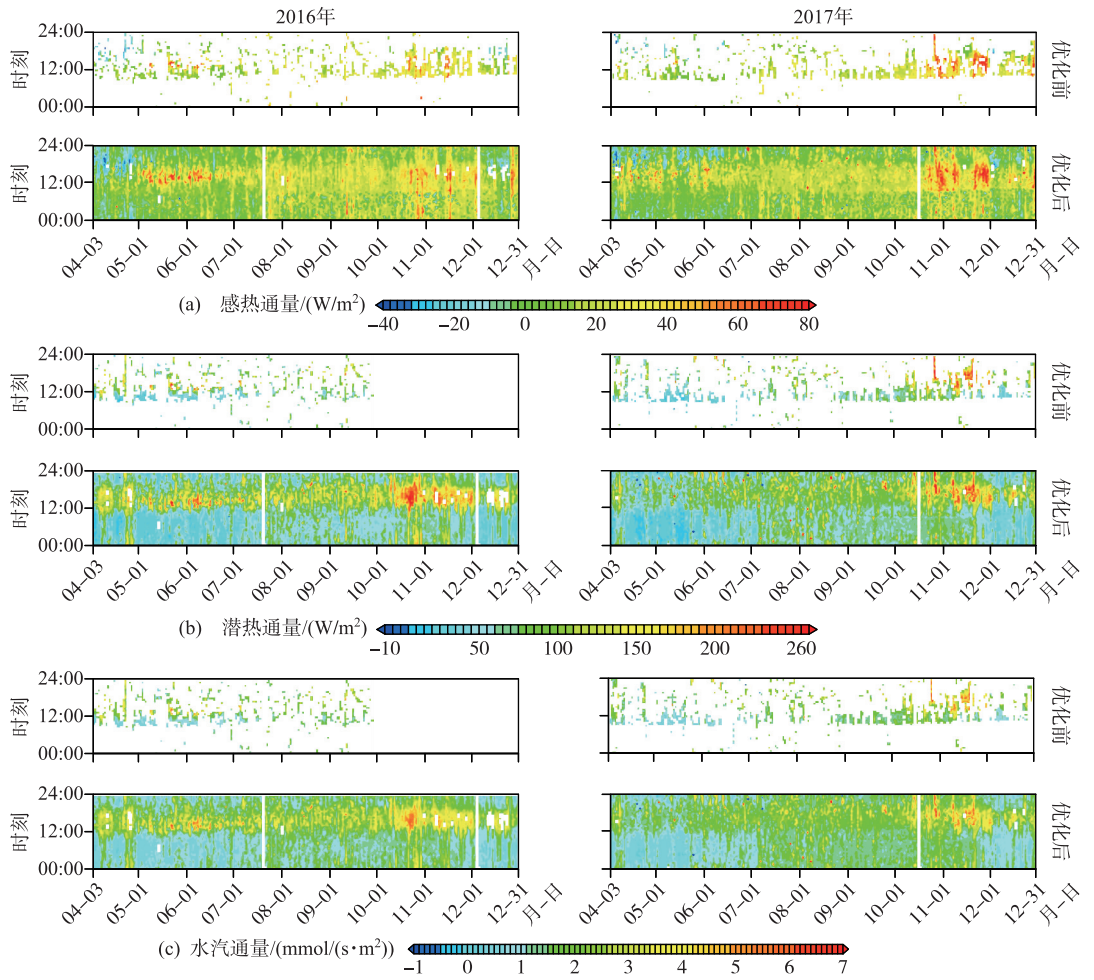

图 5 湖面通量插补前后对比:(a)感热通量,色标分辨率为 2 W/m²;(b)潜热通量,色标分辨率为 5 W/m²;(c)水汽通量,色标分辨率为 0.1 mmol/(s·m²)(时间范围为 2016 年和 2017 年的 4 月 3 日—12 月 31 日)

Fig.5 Comparison between the unpatched and patched lake surface fluxes: (a) sensible heat, with a palette resolution of 2 W/m²; (b) latent heat, with a palette resolution of 5 W/m²; (c) molar water vapor, with a palette resolution of 0.1 mmol/(s·m²) (all included in the period of Apr. 3 and Dec. 31 in 2016 and 2017)

充足和模型训练的时效,是 ANN 模型在通量数据插补问题上泛化的必要条件. 本研究对 ANN 模拟通量的验证分析表明,超宽 ANN 结构饱和输入同步观测要素的方式利用了更多有效信息. 本文将并行计算技术作为辅助工具引入通量观测研究之中,这显然是湍流通量观测研究所需进一步发展的方向. 基于大数据理论,通量模拟在数据量更大的区间效果更好,表明随着数据量增加通量模拟的效果还具备提升潜力. 本文验证 ANN 模型性能时发现,同步观测特征与通量强度通过超宽结构 ANN 模型完成了较好的映射关系拟合,这实现了同步观测特征变量自相关信息的有效利用. 因此,在今后的 EC 湍流通量观测实验中,通过增加同步观测要素,可以利用此插补方法改善高寒湖泊环境所带来的通量数据质量问题. 本文提出基于同步观测信息充分利用的数据插补思路为高寒湖泊等特殊环境下的 EC 通量观测实验提供了提升数据质量的参考.

4 参考文献

[1] Blanken PD, Rouse WR, Culf AD *et al*. Eddy covariance measurements of evaporation from Great Slave lake, Northwest

Territories, Canada. *Water Resources Research*, 2000, **36**(4): 1069-1077. DOI: 10.1029/1999WR900338.

[2] Biermann T, Babel W, Ma WQ *et al*. Turbulent flux observations and modelling over a shallow lake and a wet grassland in the Nam Co basin, Tibetan Plateau. *Theoretical and Applied Climatology*, 2014, **116**(1/2): 301-316. DOI: 10.1007/s00704-013-0953-6.

[3] Lee XH, Liu SD, Xiao W *et al*. The Taihu eddy flux network: an observational program on energy, water, and greenhouse gas fluxes of a large freshwater lake. *Bulletin of the American Meteorological Society*, 2014, **95**(10): 1583-1594. DOI: 10.1175/BAMS-D-13-00136.1.

[4] McGloin R, McGowan H, McJannet D *et al*. Quantification of surface energy fluxes from a small water body using scintillometry and eddy covariance. *Water Resources Research*, 2014, **50**(1): 494-513. DOI: 10.1002/2013WR013899.

[5] Zhao XS, Liu YB. Phase transition of surface energy exchange in China's largest freshwater lake. *Agricultural and Forest Meteorology*, 2017, **244**: 98-110. DOI: 10.1016/j.agrformet.2017.05.024.

[6] Zhao XS, Liu YB. Variability of surface heat fluxes and its driving forces at different time scales over a large ephemeral lake in China. *Journal of Geophysical Research: Atmospheres*, 2018, **123**(10): 4939-4957. DOI: 10.1029/2017JD027437.

[7] Liu HZ, Feng JW, Sun J *et al*. Eddy covariance measurements of water vapor and CO_2 fluxes above the Erhai Lake. *Science China Earth Sciences*, 2015, **58**(3): 317-322. DOI: 10.1007/s11430-014-4828-1.

[8] Wang BB, Ma YM, Chen XL *et al*. Observation and simulation of lake-air heat and water transfer processes in a high-altitude shallow lake on the Tibetan Plateau. *Journal of Geophysical Research: Atmospheres*, 2015, **120**(24): 12327-12344. DOI: 10.1002/2015JD023863.

[9] Soloway AD, Amiro BD, Dunn AL *et al*. Carbon neutral or a sink? Uncertainty caused by gap-filling long-term flux measurements for an old-growth boreal black spruce forest. *Agricultural and Forest Meteorology*, 2017, **233**: 110-121. DOI: 10.1016/j.agrformet.2016.11.005.

[10] Aubinet M, Vesala T, Papale D *et al* eds. Eddy covariance: a practical guide to measurement and data analysis. Springer Science & Business Media, 2012: 85-101. DOI: 10.1007/978-94-007-2351-1.

[11] Liu HZ, Feng JW, Sun JH *et al*. Eddy covariance measurements of water vapor and CO_2 fluxes above the Erhai Lake. *Science China: Earth Sciences*, 2014, **44**(11): 2527-2539. DOI: 10.1007/s11430-014-4828-1. [刘辉志, 冯健武, 孙绩华等. 洱海湖气界面水汽和二氧化碳通量交换特征. 中国科学:地球科学, 2014, **44**(11): 2527-2539.

[12] Li ZG, Lyu SH, Ao YH *et al*. Long-term energy flux and radiation balance observations over Lake Ngoring, Tibetan Plateau. *Atmospheric Research*, 2015, **155**: 13-25. DOI: 10.1016/j.atmosres.2014.11.019.

[13] Wang BB, Ma YM, Ma WQ *et al*. Physical controls on half-hourly, daily, and monthly turbulent flux and energy budget over a high-altitude small lake on the Tibetan Plateau. *Journal of Geophysical Research: Atmospheres*, 2017, **122**(4): 2289-2303. DOI: 10.1002/2016JD026109.

[14] Tanentzap AJ, Hamilton DP, Yan ND. Calibrating the Dynamic Reservoir Simulation Model (DYRESM) and filling required data gaps for one-dimensional thermal profile predictions in a boreal lake. *Limnology and Oceanography: Methods*, 2007, **5**(12): 484-494. DOI: 10.4319/lom.2007.5.484.

[15] Falge E, Baldocchi D, Olson R *et al*. Gap filling strategies for long term energy flux data sets. *Agricultural and Forest Meteorology*, 2001, **107**(1): 71-77. DOI: 10.1016/S0168-1923(00)00235-5.

[16] Alavi N, Warland JS, Berg AA. Filling gaps in evapotranspiration measurements for water budget studies: evaluation of a Kalman filtering approach. *Agricultural and Forest Meteorology*, 2006, **141**(1): 57-66. DOI: 10.1016/j.agrformet.2006.09.011.

[17] Chen YY, Chu CR, Li MH. A gap-filling model for eddy covariance latent heat flux: Estimating evapotranspiration of a subtropical seasonal evergreen broad-leaved forest as an example. *Journal of Hydrology*, 2012, **468**: 101-110. DOI: 10.1016/j.jhydrol.2012.08.026.

[18] Shen PK, Zhang XQ. Analysis on the observation of atmospheric turbulence characteristics over the Yamzhog Yumco, South Tibet. *J Lake Sci*, 2019, **31**(1): 243-255. DOI: 10.18307/2019.0123. [沈鹏珂, 张雪芹. 藏南羊卓雍错湖面大气湍流特征观测分析. 湖泊科学, 2019, **31**(1): 243-255.]

[19] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*, 2015, **349**(6245): 255-260. DOI: 10.1126/science.aaa8415.

[20] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Cognitive Modeling*, 1988, (3): 1.

[21] Lee JH, Xiao LC, Schoenholz SS *et al*. Wide neural networks of any depth evolve as linear models under gradient descent. arXiv preprint, 2019, arXiv:1902.06720. DOI: 1902.06720.

[22] Chu D, Pu Q, Laba Z *et al*. Remote sensing analysis on lake area variations of Yamzho Yumco in Tibetan Plateau over the past 40 a. *J Lake Sci*, 2012, **24**(3): 494-502. DOI: 10.18307/2012.0324. [除多, 普穷, 拉巴卓玛等. 近 40a 西藏羊卓雍错湖泊面积变化遥感分析. 湖泊科学, 2012, **24**(3): 494-502.]

[23] Guyon I, Gunn S, Nikravesh M *et al* eds. Feature extraction: foundations and applications. New York: Springer, 2008.

[24] Foken T. 50 years of the Monin-Obukhov similarity theory. *Boundary-Layer Meteorology*, 2006, **119**(3): 431-447. DOI: 10.1007/s10546-006-9048-6.

[25] Specht DF. A general regression neural network. *IEEE Transactions on Neural Networks*, 1991, **2**(6): 568-576. DOI: 10.1109/72.97934.

[26] Abadi M, TensorFlow AA. Large-scale machine learning on heterogeneous distributed systems. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, USA, 2016: 265-283.

[27] Chollet F ed. Keras: The python deep learning library. Astrophysics Source Code Library, 2018.

[28] Bottou L. Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT, 2010: 177-186. DOI: 10.1007/978-3-7908-2604-3_16.

[29] Srivastava N, Hinton G, Krizhevsky A *et al*. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 2014, **15**(1): 1929-1958.

[30] He KM, Zhang XG, Ren SQ *et al*. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1026-1034.

[31] Fahlman SE, Lebiere C. The cascade-correlation learning architecture. Advances in Neural Information Processing Systems, 1990: 524-532.

[32] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, 1995, **14**(2): 1137-1145.

[33] Stull RB ed. An introduction to boundary layer meteorology. Springer Science & Business Media, 2012: 14-21.

[34] Neumeyer N, Dette H, Nagel ER. A note on testing symmetry of the error distribution in linear regression models. *Nonparametric Statistics*, 2005, **17**(6): 697-715. DOI: 10.1080/10485250500095660.

[35] Makridakis S. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 1993, **9**(4): 527-529. DOI: 10.1016/0169-2070(93)90079-3.

[36] Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 1989, **45**(1): 255-268. DOI: 10.2307/2532051.