研究论文

DOI：10.11949/0438-1157.20260277

面向化工过程的系统性数据治理与预测研究

许文乐[1]，赵祺铭[1]，柳春[2]，黄新烨[2]，齐楚[2]，邱彤[1]

（[1]清华大学化学工程系，北京 100084；[2]中化数智科技有限公司，北京 100032）

摘要： 针对复杂化工过程数据普遍存在的强非线性、多变量强耦合及高频噪声等挑战，本文构建了一套系统性的数据治理与关键变量预测框架。该框架首先确立了规范化的数据治理机制，通过“中位绝对偏差+K近邻”的联合策略实现局部与全局异常值识别，并利用卡尔曼滤波在平滑降噪的同时最大程度保留工艺动态特征。在变量筛选与特征重构阶段，研究引入相关性分析、收敛交叉映射因果推断与方差膨胀因子分析方法，从统计关联与工艺机理双重维度剔除冗余特征并锁定关键驱动变量，同时注入物理规则特征以显著增强模型的可解释性。以某低温洗涤塔压降预测为例的实证研究表明，该框架结合TimeXer模型展现出卓越的性能，尤其在长步长预测任务中实现了更高的精度与鲁棒性，为化工过程的智能监测与前瞻性调控提供了可靠的技术范式。

关键词： 智能化建模；数据治理；相关性分析；因果推断；时序预测

中图分类号： TH 3　　**文献标志码：** A

文章编号： 0438-1157（2026）10-6064-12

Systematic data governance and forecasting research for chemical processes

XU Wenle[1], ZHAO Qiming[1], LIU Chun[2], HUANG Xinye[2], QI Chu[2], QIU Tong[1]

([1]*Department of Chemical Engineering, Tsinghua University, Beijing* 100084, *China*; [2] *Sinochem Digital Intelligence Technology Co., Ltd., Beijing* 100032, *China*)

Abstract: To address the challenges of strong nonlinearity, multivariable coupling, and high-frequency noise inherent in complex chemical process data, this paper proposes a systematic data governance and forecasting framework for key variables. The framework first establishes a standardized data governance mechanism, utilizing a joint strategy of median absolute deviation and K-nearest neighbors to precisely identify local and global anomalies. Kalman filtering is then applied to achieve smooth denoising while preserving dynamic process characteristics to the maximum extent. In the variable selection and feature reconstruction stage, correlation analysis, convergent cross-mapping causal inference, and variance inflation factor analysis are introduced to eliminate redundant features and lock in key driving variables from both statistical correlation and process mechanism dimensions. Simultaneously, physical regularity features are injected to significantly enhance the interpretability of the model. Empirical research on pressure drop prediction for a cryogenicscrubber demonstrates that the framework, integrated with the TimeXer model, exhibits superior performance. It achieves higher accuracy and robustness, particularly in long-term forecasting tasks, providing a reliable technical paradigm for intelligent monitoring and proactive control of chemical processes.

Keywords: intelligent modeling; data governance; correlation analysis; causal inference; time-series forecasting

收稿日期： 2026-03-03　　**修回日期：** 2026-04-19

通信作者： 邱彤(1971—)，女，教授，qiutong@tsinghua.edu.cn

第一作者： 许文乐(2001—)，男，博士研究生，xwl23@mails.tsinghua.edu.cn

引用本文： 许文乐, 赵祺铭, 柳春, 黄新烨, 齐楚, 邱彤. 面向化工过程的系统性数据治理与预测研究[J]. 化工学报, 2026, 77(10): 6064-6075

Citation: XU Wenle, ZHAO Qiming, LIU Chun, HUANG Xinye, QI Chu, QIU Tong. Systematic data governance and forecasting research for chemical processes[J]. CIESC Journal, 2026, 77(10): 6064-6075

引 言

当前，以人工智能、大数据为核心的数字技术正重塑全球工业格局，推动化工行业培育“新质生产力”并实现高质量发展[1]。深入挖掘海量工业数据中蕴含的工艺机理与演变规律，已成为化工数智化转型的核心逻辑。然而，化工装置普遍具有强非线性、多变量耦合及大时滞特性，传统机理建模难以兼顾计算实时性与复杂工况的全覆盖。相比之下，数据驱动方法能够自动提取高维动态关联，为复杂系统的监测与调控提供了全新范式[2]。

时序预测作为实现化工装置前瞻性调控、风险预警及难测变量软测量的核心技术，具有重要的工程价值。近年来，基于深度学习的数据驱动方法在化工时序预测任务中展现出卓越性能，Bai等[3]提出了一种基于Transformer的多变量多步滚动预测方法，该方法在连续搅拌罐加热器（continuous stirred tank heater，CSTH）系统和田纳西-伊斯曼过程（Tennessee Eastman process，TEP）的测试中展现出优于传统统计方法的预测精度与故障预警能力。该团队进一步提出了一种能够量化两个序列之间的形状差异的可微损失函数ESE，并通过MSE和ESE为损失函数分两步训练深度学习模型，能够有效获取化工过程数据的趋势和形状特征。针对长序列预测的参数调优难题，研究者引入了BOHB（贝叶斯优化与Hyperband算法）优化后的Informer模型，显著提升了油田产量预测等复杂场景的建模效率[4]。为了克服工业数据中的非平稳性问题，多尺度杂交分解特征提取网被提出用于工业过程的长步长预测，通过对特征进行多尺度分解，增强了模型在长期预测任务中的鲁棒性[5]。针对化工系统固有的强耦合特性，基于深度图学习（deep graph learning）的多元聚类方法被用于挖掘变量间的拓扑关联，为复杂系统的时序演变提供了更深层次的表征[6]。此外，Dong等[7]提出了一种结合时序卷积网络（temporal convolutional network，TCN）与Transformer的深度神经网络，并通过随机森林进行变量选择，能够实现对合成氨过程关键输出变量的动态建模与高精度预测。

然而，化工时序预测并非单纯的算法演进，而是一项要求数据特性、预测目标与工艺机理深度协同的系统工程。现有研究多聚焦于预测算法的创新，忽视了从原始数据流向预测应用全流程中的系统性治理与机理嵌入。事实上，受传感器漂移及环境干扰影响，工业原始时序数据普遍存在高频噪声、异常值、非随机性缺失及严重冗余等质量瓶颈。若缺乏规范的数据治理机制，端到端的“黑盒”模型在处理实际复杂工况时，其预测精度与泛化鲁棒性将难以得到根本保障。

针对上述挑战，本文提出了一套面向复杂化工过程关键变量的系统性数据驱动预测框架。该框架构建了“中位绝对偏差+K近邻”联合异常值处理策略，在保留工艺动态特征的同时实现了局部敏感与全局稳健的协同识别；通过低时滞卡尔曼滤波的“预测-更新”迭代过程，有效提升了复杂化工信号的降噪效果。在输入空间优化阶段，本文整合相关性分析、收敛交叉映射因果推断与方差膨胀因子分析，从统计关联与工艺机理维度确定关键驱动变量并剔除多重共线性冗余。最终，借助TimeXer架构对目标变量自相关性与协变量时滞耦合的精细建模，提升了复杂化工系统在时序预测任务中的精度与鲁棒性，为破解工业数据质量瓶颈、实现高精度状态感知提供了可靠的技术范式。

1 数据驱动框架

鉴于化工过程时序数据普遍存在的高度非线性、多变量强耦合、严重共线性及显著时滞等复杂特性，直接应用深度学习算法往往难以有效捕捉系统的本质动态规律，限制了预测模型的精度与泛化能力。为此，本文构建了一套面向复杂化工过程关键变量的系统性数据驱动预测框架，包含了规范化的数据治理与特征重构，以及针对化工过程的模型选择与构建。如图1所示，该框架包含时序数据采集、时序数据预处理、变量筛选与特征工程、以及时序预测模型构建四个核心阶段。

1.1 时序数据采集

时序数据采集是数据驱动建模的基础，数据的质量直接影响预测模型的最终性能。本阶段的主要任务是明确预测目标，并结合过程机理与专家经验确定输入变量集。首先，依据工艺管道仪表流程图（P&ID）等工艺资料，筛选与目标变量存在物料或能量关联的上下游变量或参数；进一步地，考虑现场操作人员的经验，将被认为对预测目标有影响的关键控制变量纳入采集范围。在确定变量集后，通过分布式控制系统（DCS）的标准接口获取历史数

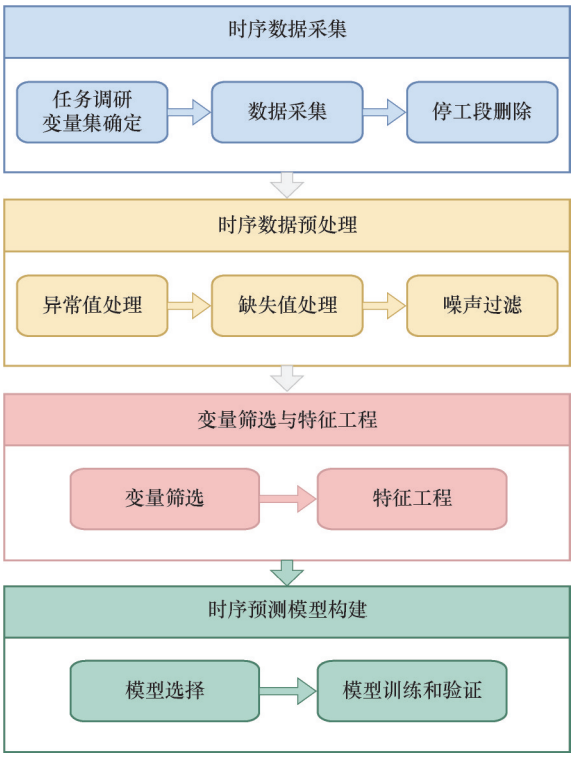

图1 针对化工过程变量时序预测的数据驱动框架

Fig.1 Data-driven framework for time-series prediction of chemical process

据。在此基础上，为剔除非正常工况（如停车阶段）下的无效数据，选取主物料流量或关键操作流量作为判别指标，当其低于预设阈值时，将对应时间段判定为非稳定生产状态并予以剔除，该阈值通常需要结合具体工艺特性以及工程师和现场操作人员的知识进行确定。以上步骤旨在构建具有统计代表性的时序数据集，为后续的预处理和建模提供可靠的数据输入。

1.2 时序数据预处理

1.2.1 异常值处理 受传感器故障、仪表漂移及数据通信异常等因素影响，采集的数据中常包含偏离装置稳态或正常过渡态工况的极端值，即异常值。直接使用包含异常值的数据进行分析，可能改变数据集的统计特性，进而影响时序预测模型的训练效果及泛化能力。因此，在特征工程与建模前，对异常值进行识别与处理是保障模型性能的必要步骤。

异常值处理流程通常包括检测、剔除与填补三个环节。如表1所示，根据输入数据的维度可将现有的时序数据异常检测方法进行分类[8]。其中，单变量时序数据仅包含单一变量的时间序列，而多元时序数据则由同一时间戳下的多变量向量构成。对于多元时序数据，既可采用建模变量间相关性的多元检测方法，亦可将其分解为多个维度，独立应用单变量检测方法。

在单变量检测领域，基于模型估计和预测的中位绝对偏差（median absolute deviation，MAD）方法对极端值具有较强的鲁棒性，常用于滑动窗口内的局部稳健估计；基于邻域密度的局部异常因子（local outlier factor，LOF）方法通过对比样本点与其邻域点的局部可达密度，能够有效识别分布不均匀数据集中的局部异常。而基于直方图的HBOS算法凭借极低的计算复杂度，在处理高频采集的大规模工业实时数据流方面极具优势。相比之下，多变量检测侧重于捕捉系统整体状态的关联性异常，其技术路径涵盖了从高维空间向低维子空间的降维转换、将标量统计量扩展为向量范数的直接建模，以及利用图模型或聚类算法对高维流形空间几何结构的深度挖掘，从而识别在单一变量维度下难以察觉的系统性偏离。

在识别到异常点后，通用做法是先从原始数据中对异常点进行剔除，并通过缺失值处理进行填补，具体方法将在后续章节进行介绍。

表1 多元时序数据异常值检测方法

Table 1 Outlier detection methods for multivariate timeseries

输入数据	方法类别	代表方法	核心思想
单变量	基于模型估计和预测	中位绝对偏差（MAD）、ARIMA	利用历史及未来数据拟合模型，计算观测值与估计值的残差 $\lvert x_t - \hat{x}_t \rvert$ 作为判据
	基于邻域密度	局部异常因子（LOF）、K-近邻（KNN）	基于数据点的邻域密度判断是否属于异常值
	基于直方图	基于直方图的离群点检测（HBOS）	将样本分为多个区间，样本数少的区间内的数据点被判定为异常值
多变量	降维与单变量转换	主成分分析（PCA）、自动编码器 Autoencoder	将多元高维数据映射到低维空间或转换为单变量序列，再应用单变量检测技术
	单变量方法扩展	VAR、基于高维距离的KNN和LOF	将单变量的标量运算推广为向量运算，直接对多维数据建模或统计，如多元模型的估计和预测等方法
	基于相异度与几何结构	基于图结构的检测算法[9]	直接构建数据点在空间中的几何或拓扑关系

1.2.2 缺失值处理 原始时间序列中不可避免地存在数据缺失现象，若不进行有效处理，不连续的数据序列将破坏时间序列的内在连续性与动态依赖关系，进而干扰预测模型的训练。针对不同尺度的缺失模式，处理策略也会有差异。对于少量的随机散发性缺失，通常采用向前填充、统计填充（如均值、中位数）或插值法（线性插值、样条插值），以在维持过程数据惯性特征的同时，保证序列的局部平滑性；对于长时段的连续缺失，理论上可采用基于模型的方法（如KNN、随机森林）或序列预测模型（如ARIMA）进行重建。然而，针对时序预测任务的数据清洗环节，对长片段缺失的强行填补往往会导致数据波动模式的过度平滑化或失真，从而误导模型学习到简化的虚假规律。

1.2.3 噪声过滤 对于实际化工过程，原始采集数据中常伴有高频噪声，为提取真实的工艺动态特征，需对数据进行平滑降噪处理。值得注意的是，针对时序预测任务，降噪过程必须严格遵循“因果性约束”，即对于t时刻的数据处理严禁依赖t+1及后续时刻的数据，违反此原则将导致离线测试指标失真，无法客观反映模型的在线预测性能。常用的因果降噪算法主要包括时域统计类、多项式拟合类及状态空间类方法。

对于时域统计类方法，选择简单移动平均通过计算滑动窗口内的算术均值滤除高频噪声，但会引入较高的滞后；指数加权移动平均（EWMA）通过赋予近期数据更高权重，提升了对趋势变化的响应速度；因果中值滤波则利用窗口内的中位数进行稳健估计，在去除脉冲噪声的同时有效保持了信号的阶跃边缘。对于多项式拟合方法，代表算法为因果Savitzky-Golay滤波器。该方法通过最小二乘法在滑动窗口内拟合低阶多项式，相较于简单的时域统计方法，其在去噪的同时能更有效地保留波峰、波谷等几何特征[10]。对于状态空间类算法，代表算法为卡尔曼滤波（Kalman filter），该算法假设系统状态由线性方程演化且过程噪声与测量噪声均服从高斯分布，通过“预测-更新”的迭代过程实现最优估计。

（1）预测阶段：基于上一时刻的最优估计$\hat{x}_{k-1}$和状态转移矩阵$\boldsymbol{F}$，预测当前时刻的状态$\hat{\boldsymbol{x}}_k^-$，并预测该估计的不确定性（协方差）$\boldsymbol{P}_k^-$，其中，$\boldsymbol{Q}$是过程噪声矩阵：

$$\hat{\boldsymbol{x}}_k^- = \boldsymbol{F}\hat{\boldsymbol{x}}_{k-1} \tag{1}$$

$$\boldsymbol{P}_k^- = \boldsymbol{F}\boldsymbol{P}_{k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} \tag{2}$$

（2）更新阶段：首先计算卡尔曼增益$\boldsymbol{K}_k$，用于权衡测量值$\boldsymbol{z}_k$（噪声协方差为$\boldsymbol{R}$）与预测值$\hat{\boldsymbol{x}}_k^-$的可信程度，进而修正先验估计得到当前时刻的最优估计$\hat{\boldsymbol{x}}_k$，并更新协方差$\boldsymbol{P}_k$为下一次迭代做准备：

$$\boldsymbol{K}_k = \boldsymbol{P}_k^-\boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{H}\boldsymbol{P}_k^-\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}\right)^{-1} \tag{3}$$

$$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k^- + \boldsymbol{K}_k\left(\boldsymbol{z}_k - \boldsymbol{H}\hat{\boldsymbol{x}}_k^-\right) \tag{4}$$

$$\boldsymbol{P}_k = \left(\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}\right)\boldsymbol{P}_k^- \tag{5}$$

基于此迭代过程，卡尔曼滤波能够敏锐跟踪数据的动态变化，在有效滤除随机噪声的同时，最大限度地保留了反映工艺本质的形态特征与趋势信息。

1.3 关键变量筛选与特征工程

尽管前序的数据清洗环节显著提升了原始信号的统计质量，但鉴于化工过程固有的高维强耦合、大时滞及非线性动力学特性，直接将全量原始测点输入模型仍面临严峻挑战。多变量间普遍存在的严重多重共线性不仅会导致大量信息冗余，引发“维数灾难”，还会显著增加模型过拟合的风险，从而掩盖变量间真实的拓扑依赖关系。因此，为构建兼具高精度与物理一致性的预测模型，必须执行严格的变量筛选与特征工程。一方面，通过变量筛选剔除无关噪声变量与高度冗余特征，锁定对目标变量具有实质驱动作用的关键协变量子集；另一方面，借助特征工程将工艺机理与专家经验（如无量纲准数、特定的非线性组合）显式编码为模型可识别的数学特征，向纯数据驱动模型注入物理约束，增强模型在复杂工况下的鲁棒性与泛化能力。

1.3.1 关键变量筛选 变量筛选的核心任务是构建兼具高预测能力与低冗余度的特征子集。

首先，通过相关性分析量化协变量与目标变量的统计依赖强度，初步剔除无关或弱相关变量。其中，皮尔逊相关系数（Pearson correlation coefficient）通过计算两个时间序列的协方差与各自标准差乘积之商，量化变量间的线性相关程度，但受限于化工过程显著的非线性特征，其在捕捉复杂工况关联方面存在局限。斯皮尔曼相关系数（Spearman correlation coefficient）通过评估变量间的单调关系，克服了对数据分布的严格假设，对异常值具有较强鲁棒性，更适用于处理含有噪声的复杂时序数据。此外，还可以通过互信息（mutual information）计算联合概率分布与边缘概率分布乘积之间的Kullback-Leibler（KL）散度来度量变量间的依赖关

系，该指标能有效识别包括复杂非线性耦合在内的任意形式统计依赖，适合捕捉化工变量间隐含的高维关联[11]。

其次，引入因果分析验证协变量对目标变量的驱动作用，该阶段旨在识别变量间的方向性依赖关系，从而排除“伪相关”并构建具有物理可解释性的特征集。格兰杰因果检验（Granger causality）基于“预测因果性”假设，通过评估引入协变量历史信息后对目标变量预测误差的降低程度来判定因果，该方法计算效率高，但主要适用于线性系统[12]。传递熵（transfer entropy）被视为非线性场景下的格兰杰因果，通过计算条件互信息量化不对称的信息流动强度，由于不依赖特定数学模型，其具备捕捉任意形式非线性耦合的能力，更适用于复杂化工过程变量分析[13]。收敛交叉映射（convergent cross mapping）基于非线性动力系统状态空间重构理论，通过“影子流形”反向预测状态，能有效区分因果关系与共同外部驱动的同步现象，特别适用于解析化工装置中带有强耦合与闭环反馈特性的复杂结构[14]。

在筛选出具有因果关系的特征子集后，变量间往往仍存在显著的多重共线性。若将高度共线的变量同时输入模型，会导致参数估计不稳、模型过拟合及可解释性下降。除了利用相关系数矩阵进行初步排查外，还可以通过方差膨胀因子（variance inflation factor, VIF）衡量多变量之间的共线性强度[15]。该方法将特征集中的某一变量X_i作为因变量，利用其余所有特征作为自变量进行线性回归，并根据回归方程的决定系数R_i^2计算VIF值，通常认为，当$\mathrm{VIF}_i = \dfrac{1}{1 - R_i^2} > 10$时，表明该变量与其他变量间存在严重的共线性，应予以剔除。

1.3.2 特征工程 特征工程是将原始DCS数据转化为模型可理解的高效输入特征的关键环节。尽管LSTM或Transformer等深度学习模型具备强大的自动表征学习能力，在一定程度上降低了对人工特征的依赖，但有针对性的特征构建仍不可或缺。通过引入基于机理的特征（如差分计算捕捉变化速率、滞后特征显式表达因果时延），实质上是将领域知识与物理规律注入数据驱动模型。

常用的特征构建方法可归纳为基于统计的方法与基于机理的方法两类。其中，基于统计的特征侧重于挖掘数据的时间序列属性，包括利用一阶或二阶差分特征捕捉变量的瞬态变化速度及加速度；应用信号滤波特征分离长期趋势与高频噪声；以及计算滑动窗口统计量（如均值、方差、极值），以量化局部的基准趋势、运行边界及波动幅度。而基于机理的特征构建侧重于融合过程先验知识以构建符合物理约束的复合特征，如构建差值和比值特征描述某股流体的进入换热器前后的温度变化、构建质量和能量衡算特征显式描述设备的物理运行状态，以及针对强耦合变量设计非线性组合（如通过流量与温差的乘积构造热负荷指标），以显式表达变量间的交互作用。

1.4 时间序列预测

时序预测的核心任务在于通过建模历史观测数据中的趋势性、周期性、自相关性及变量间的互相关性，实现对目标变量未来状态的推演。根据技术路线的差异，现有方法主要分为统计学习方法与深度学习方法。

首先是统计学习方法，侧重于对数据生成过程建立线性统计模型。自回归模型（autoregressive model, AR）通过目标变量历史值的线性组合预测未来走势，能够捕捉序列自身的线性依赖，但难以处理非线性动态及多变量耦合；差分整合自回归移动平均模型（autoregressive integrated moving average, ARIMA）在AR基础上引入差分操作以解决序列非平稳问题，并结合移动平均模块平滑预测误差，是应用最广的统计模型之一，但其本质仍受限于线性模型体系，在面对高度非线性的化工过程时建模能力存在局限。

其次是深度学习方法，凭借强大的表征学习能力，已成为处理化工高维非线性时序数据的主流技术。卷积神经网络（convolutional neural network, CNN）通过卷积核的局部感应与权值共享机制，实现对时序模式的分层提取，擅长捕捉序列的局部相关性并降低模型复杂度；循环神经网络（recurrent neural network, RNN）及其变体（如LSTM、GRU）通过递归机制和精巧的门控单元构建“记忆”功能，有效解决了基础神经网络难以处理长程依赖及梯度消失的问题，适合建模具有时滞效应的化工动态过程；Transformer架构则摒弃了传统的顺序约束，利用自注意力机制（self-attention）直接计算序列内任意位置间的依赖权重，不仅具备卓越的全局信息整合能力，且支持并行计算，显著提升了大规模长时序数据的训练效率与预测精度。

然而，在处理化工过程典型的多变量时间序列

时，经典的Transformer架构存在一定局限性。传统方法通常利用简单的嵌入层将同一时间步的所有变量数值映射为一个统一的时间词元。这种编码方式在早期阶段即混合了不同变量的信息，容易导致多变量间复杂耦合关系的过度简化，且难以有效捕捉变量间普遍存在的时滞特性，从而限制了模型在多输入变量场景下的预测性能。通过模型结构的改进可以在一定程度上解决这个问题，以TimeXer为例，其对协变量和目标变量设计了差异化的嵌入策略和注意力机制[16]。对于协变量，TimeXer以变量的时间窗口为基本单位进行嵌入；而对于目标变量，则将输入的时间窗口划分为一系列不重叠的片段，并将每个片段投影为输入Token的一个点。随后，模型对目标变量的输入Token通过自注意力机制建模其时间维度上的相关性，并通过与协变量输入Token的交叉注意力机制建模不同变量之间的相关关系。这种设计不仅更精准地建模了化工场景中目标变量的自相关性及变量间的时滞耦合关系，其注意力权重分布还能直观揭示不同变量及时间片段对预测结果的贡献程度，从而显著增强了模型的可解释性。

2 案例分析

为验证本文所提系统性数据驱动框架的有效性与普适性，选取某无机合成装置中的低温洗涤塔单元作为研究对象，开展关键过程变量的时序预测应用研究。案例数据集采集自工厂的DCS，涵盖了该工段连续运行近四个月（采样间隔为1 min）的历史工况数据。本案例聚焦于低温洗涤塔压降的高精度时序预测建模任务，作为低温洗涤塔的核心参数，对低温洗涤塔压降的准确预测对于维持塔内气液平衡、预防液泛等异常工况以及实现装置的前瞻性优化调控具有显著的工程价值。

2.1 时序数据采集

在数据准备阶段，按照前文所述的数据采集流程，首先依据工艺管道仪表流程图溯源与目标变量存在物理关联的上下游参数；其次，结合现场工艺工程师的操作经验，从DCS系统中初步筛选出包括原料气流量、各流股进出各单元的温度、各段压力在内的68个过程变量作为候选输入特征。此外，依据操作日志记录与人工核查，识别并剔除了装置停车检修、启停过渡等非稳态时段的数据，从而构建了具备统计代表性的时序数据集。

2.2 时序数据预处理

在初始的数据预处理环节，鉴于变量筛选尚未实施，原始候选数据集中仍保留了68个过程变量。由于其中不可避免地混杂着与预测目标无关的冗余变量，若直接应用多变量异常检测算法，可能会受到维度灾难的影响。此外，非相关变量中的随机波动可能会掩盖关键变量中的真实异常模式，从而降低检测的准确性。因此，本部分优先对每个变量独立执行单变量异常值检测，以确保基础数据的统计真实性，为后续的变量筛选与建模奠定质量基础。

本环节系统评估了基于模型估计（MAD）、基于邻域密度（KNN）以及基于直方图（HBOS）三类代表性算法的检测效能。在实施过程中，采用可视化分析与异常比例统计相结合的策略，对判别阈值及超参数进行了多轮迭代寻优。具体而言，对于MAD方法，引入长度为120的滑动窗口以增强对局部非平稳数据的适应性。而针对KNN与HBOS这类利用全局分布信息的算法，为严格杜绝数据泄露，本研究严格遵循时序划分原则：仅利用训练集数据完成K近邻模型的构建及直方图边界的确定，随后将固定的模型参数与阈值迁移应用于验证集与测试集。

在完成了超参数的调整后，以低温洗涤塔压降变量为例，三种方法的异常值的检出比例分别为0.62%、0.91%和0.85%。图2展示了某周的整体检测效果，对比分析发现，结合滑动窗口的MAD方法对局部异常波动具有更高的敏感度，但受限于窗口机制，当前序异常值进入窗口时可能抬高局部中位数，导致后序异常出现“被掩盖”的漏检现象。相比之下，KNN和HBOS方法倾向于捕捉全局统计分布层面的显著偏离，虽能有效识别极端离群点，但往往难以敏锐感知局部工况微小的动态突变。

鉴于上述单一检测算法在局部敏感度与全局稳健性上存在显著的互补性，本研究最终确立了“MAD +KNN”的联合检测策略。如图2（d）所示，这种组合方法有效克服了单一模型的检测缺陷，能够精准识别KNN遗漏的局部瞬态波动，同时填补了MAD在窗口平滑效应下可能丢失的全局异常点。在完成异常值的识别后，本研究遵循剔除加填补的标准清洗范式，首先将判定为异常的数据点予以剔除，并将其作为缺失值进行后续处理。为杜绝数据泄露风险，在填补环节仅采用向前填充方法。

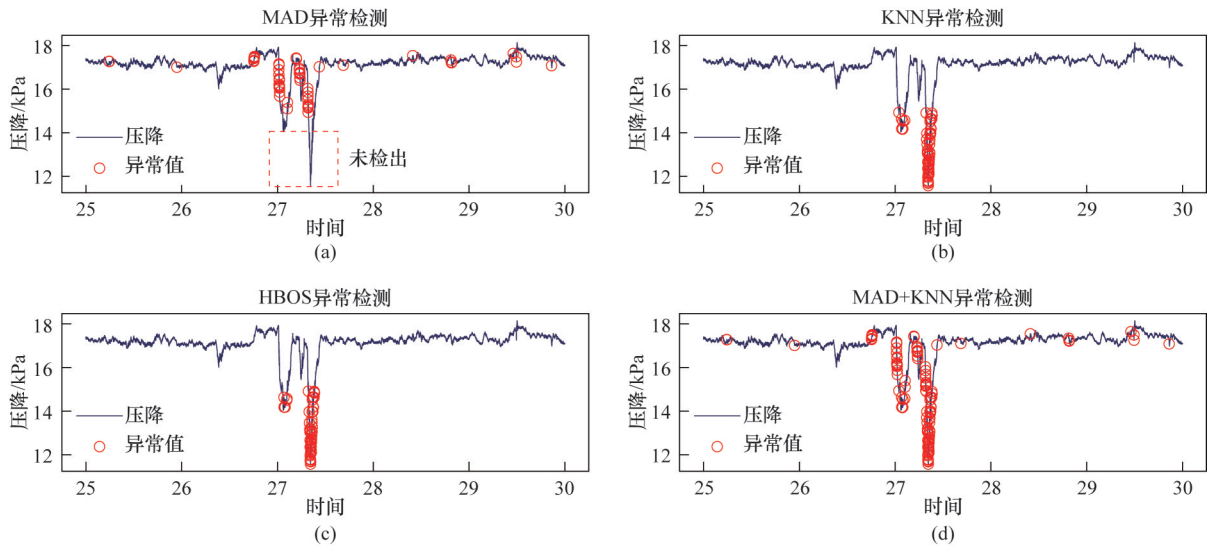

图2 异常值检测整体效果

Fig.2 Overall result of outliers detection

在完成了异常值的处理和缺失值的填补后，在数据预处理部分还需要对数据的噪声进行过滤，以获得更加平滑的用于建模的数据。对比了简单移动平均、指数加权移动平均、中值滤波、Savitzky-Golay滤波和卡尔曼滤波五种方法在本案例中的效果(图3)。简单移动平均与指数加权移动平均虽然能有效滤除高频噪声，但二者在数据发生较大变化处均存在明显的滞后性；中值滤波虽具备极强的抗离群点能力，但其平滑曲线呈现出明显的“阶梯状”特征，导致变量的局部细节丢失；Savitzky-Golay滤波方法能够保留更多的形状信息，但是其同样引入了一定的时间滞后；而相比之下，卡尔曼滤波综合表现最佳，它不仅在平稳段保持了良好的平滑度，更在07:30后的急剧下降阶段展现出较低的滞后性

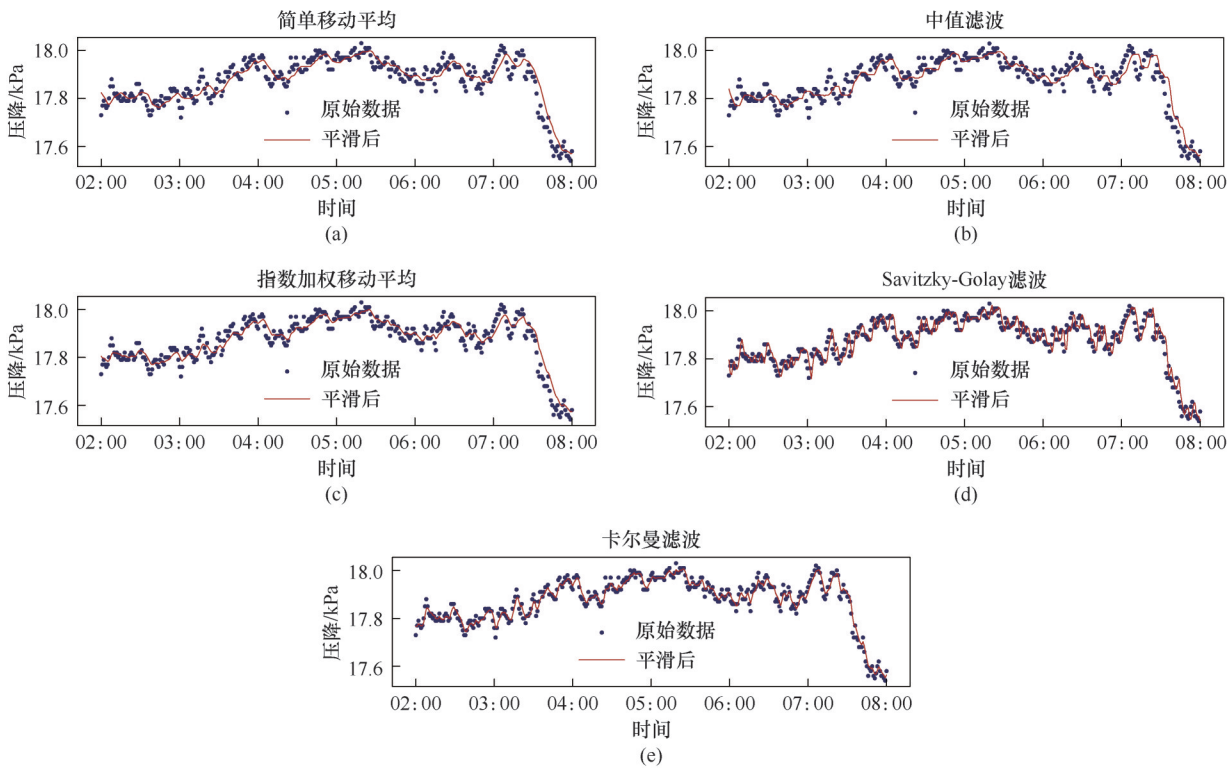

图3 噪声过滤效果展示

Fig.3 Demonstration of noise filtering effect

与较好的动态响应能力。

为了客观量化不同降噪算法在降噪性能与动态响应之间的权衡关系，本研究引入了平滑度比与滞后时间作为核心评价指标。其中，平滑度比通过计算平滑后序列的一阶差分绝对值均值与原始序列的一阶差分绝对值均值之比来获得。S_{ratio}值越小表明平滑后的曲线波动越小，平滑效果越显著，其计算公式如下：

$$S_{\text{ratio}} = \frac{\sum_{t=2}^{N} \left| x^{\text{smoothed}}{}_{t} - x^{\text{smoothed}}{}_{t-1} \right|}{\sum_{t=2}^{N} \left| x_t - x_{t-1} \right|} \tag{6}$$

滞后时间用于评估平滑算法在动态变化过程中的响应延迟，该指标通过在预设的时间窗口$\left[-\tau_{\max}, \tau_{\max} \right]$内寻找一个位移量$\tau$，使得原始序列与经过位移后的平滑序列之间的皮尔逊相关系数达到最大。该位移量τ_{opt}代表了平滑后的时间序列信号相对于原始序列的平均时间延迟，计算公式如下：

$$\tau_{\text{opt}} = \arg\max_{\tau \in} r_{\text{Pearson}}\left(\boldsymbol{X}, \boldsymbol{X}^{\text{smoothed}} \right) \tag{7}$$

量化评估结果如表2所示，简单移动平均与中值滤波虽然取得了最低的平滑度比，但付出了最大的时间代价，其滞后时间均超过4 min，且中值滤波在波形上呈现出丢失细节的“阶梯效应”。指数加权移动平均在保持同等平滑水平的基础上，将滞后时间缩短至3.04 min，改善了动态跟随性。Savitzky-Golay滤波器通过局部多项式拟合实现平滑，其特点在于能够较好地保留信号的局部几何特征，例如波峰、波谷位置、局部斜率及曲率变化等。因此，其平滑度比数值相对较高，表明其对高频波动的抑制作用较为温和。而本研究所关注的目标变量为分钟级的低温洗涤塔压降信号，这类高频几何特征往往包含较多由传感器噪声或过程扰动引入的随机波动，其保留反而可能降低模型对有效趋势的学习能力。相比之下，卡尔曼滤波展现了最优的综合性能，其略高的平滑度比实则反映了对真实物理信号细节的有效保留，且相较于其他方法，此方法的滞后时间仅为1.48 min，在去噪保真与动态响应之间取得了最佳平衡。

综上所述，通过集成异常值清洗、缺失值填补与平滑降噪的系统性预处理，显著提升了原始工业数据的信噪比与统计一致性，为后续的关键变量识别、特征空间重构及高精度时序建模构筑了坚实的数据根基。值得强调的是，数据预处理并非单一算法的机械应用，而是一个需要多方案横向比选的动态过程，必须构建科学的评价体系，以优选出与特定业务场景最适配的处理路径。

表2 平滑方法指标

Table 2 Denoising metrics

平滑方法	S_{ratio}	τ_{opt}/min
简单移动平均	0.46	4.08
指数加权移动平均	0.47	3.04
中值滤波方法	0.47	4.16
Savitzky-Golay滤波	1.01	2.01
卡尔曼滤波	0.66	1.48

2.3 变量筛选与特征工程

本环节聚焦于模型输入空间的深度优化，旨在构建精简且高效的特征子集。一方面，从统计与因果层面筛选出与预测目标存在实质关联的核心变量；另一方面，通过特征工程将专家经验与物理规则显式编码为高维数学特征。

2.3.1 变量筛选 首先，作为变量筛选的第一步，本研究采用多维度的相关性度量策略对候选变量进行初筛。为了全面捕捉线性、单调及复杂非线性关系，同时计算了皮尔逊相关系数、斯皮尔曼等级相关系数以及互信息。如图4所示，分析结果表明，变量F1、P3、P2、T22和F6在三种不同的度量指标下均表现出显著且一致的高相关性，揭示了这些变量与目标压降之间存在稳固的强耦合关系。

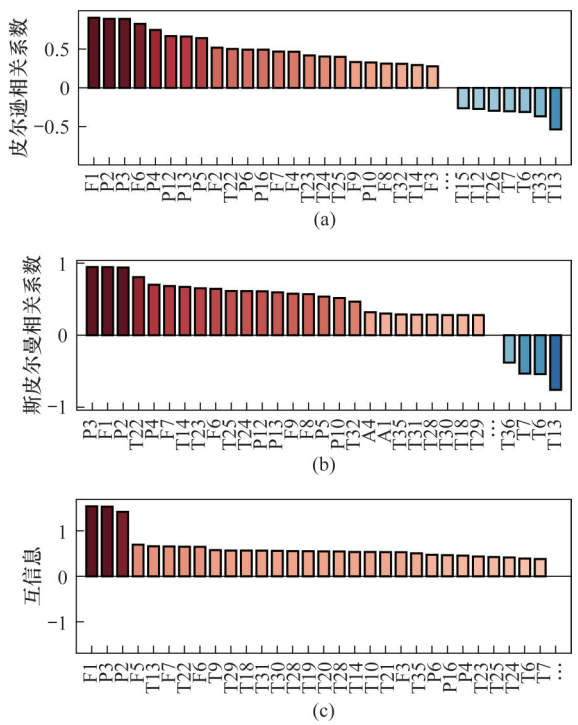

图4 变量相关性分析

Fig.4 Correlation analysis

考虑到低温洗涤塔单元内部存在复杂的换热网络与物料循环回路，且目标变量与其他过程参数间普遍存在高维非线性耦合，传统的线性因果检验往往失效。因此，本研究选用收敛交叉映射（convergent cross-mapping, CCM）算法进行因果推断。以变量F6为例（分析结果如图5所示），随着重建流形样本量的增加，双向交叉映射的预测相关系数呈现出显著的单调上升趋势，并在样本量达到10000后逐渐收敛至稳定的渐近线。这种典型的收敛特性有力证实了F6与目标变量隶属于同一个耦合动力系统，二者之间存在因果关联；同时，双向映射的高预测精度揭示了系统中存在显著的反馈机制，即变量间互为因果。基于0.6的预测能力阈值，其余对目标变量表现出强因果驱动作用的关键特征已汇总于表3。

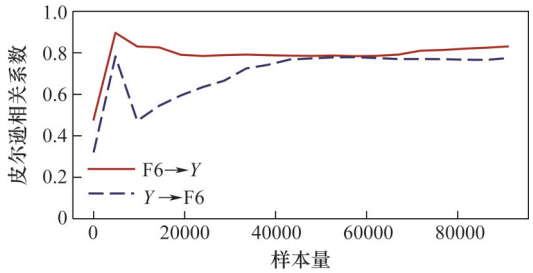

图5 CCM分析结果（F6）

Fig.5 CCM analysis results for F6

表3 收敛交叉映射分析结果

Table 3 CCM analysis results

变量	X→Y预测能力	Y→X预测能力	变量	X→Y预测能力	Y→X预测能力
F1	0.89	0.92	P5	0.67	0.72
P3	0.88	0.91	T22	0.64	0.51
P2	0.88	0.9	F2	0.63	0.45
F6	0.87	0.81	P13	0.63	0.63
P4	0.74	0.8	P12	0.62	0.64

基于前序的相关性与因果性联合筛选，本研究选择了12个对目标变量具有显著预测能力的候选特征并进行共线性诊断。无论是基于互相关系数的矩阵分析（图6）还是方差膨胀因子（variance inflation factor, VIF）（表4）计算，该候选子集内均存在严重的多重共线性，提示变量间存在高度的信息重叠。为消除冗余，可以对高VIF变量进行了迭代式的删减与筛选。经过多轮优化，最终确立的输入变量子集如表5所示，该集合在保留关键驱动信息的同时，显著降低了特征间的耦合度。

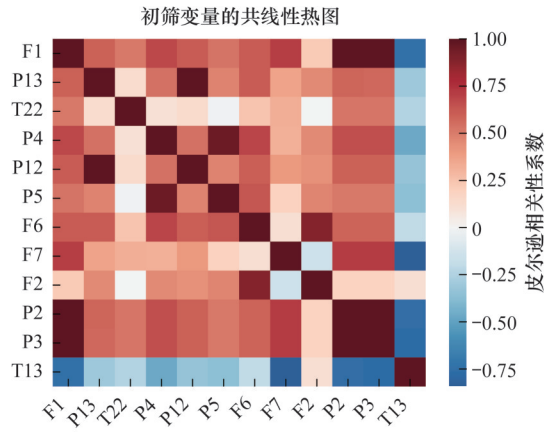

图6 初筛变量的相关性分析

Fig.6 Correlation analysis of preliminary screening variables

表4 方差膨胀因子分析

Table 4 VIF analysis

变量	VIF
P3	469.49
F1	387.24
P2	311.88
F6	109.4
F2	57.14
P12	47.42
P13	46.22
P4	23.43
P5	18.01
F7	11.03
T13	6.12

表5 变量膨胀方差因子分析（变量删减后）

Table 5 VIF analysis (after variable deletion)

变量	VIF
T23	5.49
T22	5.05
F7	4.54
T13	4.38
F6	4.13
P5	2.19
F4	2.06
P13	2.04

除上述基于统计指标的数据驱动筛选外，为了确保输入特征空间在物理层面的完备性，本研究进一步结合对低温洗涤工艺机理的理解和工程师的专家经验，将F5、T7、A4、F1及L1纳入最终的输入变量集。旨在通过专家经验的显式注入，弥补纯统计筛选可能忽略的关键过程信息，从而完善模型对装

置运行状态的表征能力。

2.3.2 特征工程 对于目标变量低温洗涤塔压降，其波动本质上是对上游工况（如进料流量）及塔内持液量变化的响应，具有一定的物理惯性与滞后特性。为此，本研究额外引入压降的一阶差分特征（即变化速率），旨在显式捕捉该变量的瞬态演变趋势，辅助模型理解系统的动态方向。与此同时，塔釜液位作为另一核心状态参数，表现为物料流入与流出速率差的积分累积过程，因此，其时序响应呈现出固有的平滑性与滞后性。引入液位的一阶差分量具有明确的物理含义，它直接映射了塔内的瞬时净物料通量状态。通过构建上述导数类特征，旨在增强输入空间对过程动态趋势与累积效应的表征能力。

2.4 时序预测

鉴于本化工过程变量间普遍存在的高维非线性耦合及目标变量复杂的自相关特性，本研究选用TimeXer作为核心预测架构，旨在深度提取时间序列内部的演变规律及变量间的交互依赖。同时评估了时间卷积网络（TCN）[17]、长短期记忆网络（LSTM）[18]及经典Transformer模型的预测效能。实验数据集严格遵循时序先后原则进行划分，设定训练集、验证集与测试集的比例为70%∶15%∶15%；在参数设置上，输入回看窗口固定为64，分别对未来8、16及32个时间步长的状态进行多步预测。训练策略方面，采用ReduceLROnPlateau动态调整学习率以适应损失函数的收敛停滞，并部署EarlyStopping早停机制防止模型过拟合。最终在独立测试集上的量化预测结果汇总于表6。

表6 时序预测指标

Table 6 Results for time-series forecasting

模型	预测长度	MSE	MAPE/%	R^2
TCN	8	1.30	4.26	−1.817
	16	3.35	7.66	−6.245
	32	25.60	20.11	−54.255
LSTM	8	1.02	5.04	−1.209
	16	2.52	8.55	−4.453
	32	2.51	8.54	−4.438
Transformer	8	0.48	2.78	−0.047
	16	0.47	2.76	−0.030
	32	0.49	2.81	−0.074
TimeXer	8	0.01	0.39	0.967
	16	0.03	0.59	0.921
	32	0.08	0.92	0.826

TimeXer模型在所有测试下均展现出显著的性能优势，相较于TCN和LSTM模型在长步长预测中出现的严重性能衰退，以及Transformer模型虽保持稳定性但精度相对有限的表现，TimeXer模型不仅在短步长下实现了较高的精度，更在长步长任务中展现了极强的鲁棒性，充分验证了该模型在捕捉化工过程长程时序依赖及多变量复杂耦合特征方面的卓越能力。从R^2结果进行分析，TimeXer在8、16及32步长下分别达到了0.967、0.921和0.826，相比之下，TCN、LSTM及Transformer模型在处理该复杂工况时的R^2普遍呈现负值，这表明这些模型无法学习到变量之间的依赖关系。在测试集上的部分预测样本如图7所示，模型输出与真实过程值之间表现出高度的一致性，并且在预测时域内的局部变化实现了紧密的跟踪，表明模型已成功学习到该变量的核心时序特征。

2.5 可解释分析

本研究提取了测试集样本在不同预测步长（8、16、32 min）下的交叉注意力权重，并进行全局均值化处理，生成了反映变量贡献度的交叉注意力热图，如图8所示。热图分析显示，F1、T22及F7在不同步长下均占据了较高的权重。这一发现与前文基于皮尔逊相关性及互信息分析的统计筛选结果高度吻合，证明了TimeXer架构能够自主捕捉化工过程中的变量耦合关系。此外，注意力分布在不同预测尺度下展现出较高的稳定性，符合化工装置的内在一致性。

3 结 论

本文针对化工过程数据的强非线性、动态时变性及高维耦合特征，进行了深度融合工艺机理与数据驱动的系统性数据治理与时序预测方法研究。研究整合了“局部敏感+全局稳健”的联合异常检测策略、基于相关性分析、因果推断与共线性诊断的层级式变量筛选机制，以及物理规则引导的特征工程方法。通过低温洗涤塔压降预测的实证研究，上述方法成功克服了单纯数据驱动方法面临的“维数灾难”与“伪相关”困境，验证了TimeXer模型在解析长程时序依赖与多变量复杂交互方面的卓越效能，为化工复杂单元的高精度状态感知提供了可复刻的技术范式。

未来研究方向将重点聚焦于模型的动态自适

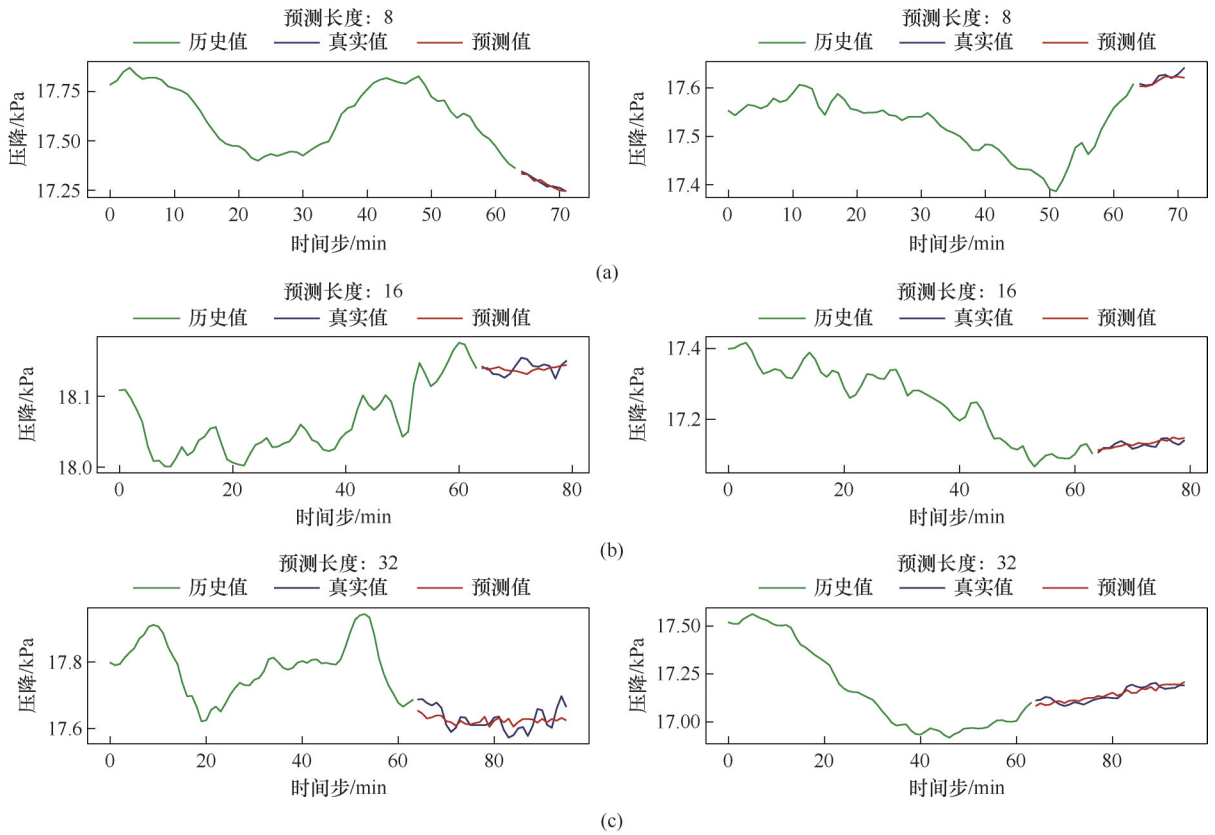

图7 时序预测结果

Fig.7 Time-series forecasting results

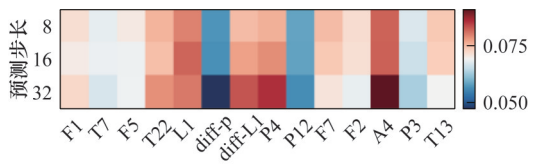

图8 模型交叉注意力热图

Fig.8 Cross attention heatmap of the models

应能力与全流程可解释性。具体包括：研发在线因果推断与增量学习算法，使模型能够实时适应装置工况漂移与设备老化带来的分布变化；探索基于图神经网络的动态拓扑建模，进一步显式解构变量间时变的耦合结构。随着先进智能算法与深层工艺规律的有机互融，预期将催生出从“经验维持”向“数智推演”跨越的生产新模式，为化工产业突破传统增长瓶颈、实现高质量数字化转型奠定坚实的科学基石。

参考文献

[1] 杨友麒. 加快发展化工新质生产力的几个问题[J]. 现代化工, 2024, **44**(12): 1-9.
Yang Y Q. Some issues in promotion of new quality productivity in chemical industry[J]. Modern Chemical Industry, 2024, **44**(12): 1-9.

[2] Ge Z Q, Song Z H, Ding S X, et al. Data mining and analytics in the process industry: the role of machine learning[J]. IEEE Access, 2017, **5**: 20590-20616.

[3] Bai Y M, Zhao J S. A novel transformer-based multi-variable multi-step prediction method for chemical process fault prognosis [J]. Process Safety and Environmental Protection, 2023, **169**: 937-947.

[4] Deng W, Xin X K, Song R X, et al. A time series forecasting method for oil production based on Informer optimized by Bayesian optimization and the hyperband algorithm (BOHB)[J]. Computers & Chemical Engineering, 2025, **197**: 109068.

[5] Wang Y, Han Y M, Hu X, et al. Long-term series forecasting for industrial processes based on multiscale hybrid decomposition feature extraction network[J]. IEEE Transactions on Industrial Informatics, 2025, **21**(7): 5401-5410.

[6] Huang J, Qin R. A novel deep graph learning-based multivariate clustering method for time series forecasting of complex chemical systems[J]. Computers & Chemical Engineering, 2025, **200**: 109214.

[7] Dong J G, Liu X N, Su R X, et al. TCN-transformer deep network with random forest for prediction of the chemical synthetic ammonia process[J]. ACS Omega, 2025, **10**(2): 2269-2279.

[8] Blázquez-García A, Conde A, Mori U, et al. A review on outlier/anomaly detection in time series data[J]. ACM Computing Surveys, 2022, **54**(3): 1-33.

[9] Cheng H, Tan P N, Potter C, et al. A robust graph-based algorithm for detection and characterization of anomalies in noisy multivariate time series[C]//IEEE International Conference on

Data Mining, 2008: 349-358.

[10] Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures[J]. Analytical Chemistry, 1964, **36**(8): 1627-1639.

[11] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information[J]. Physical Review E, 2004, **69**(6): 066138.

[12] Granger C W J. Investigating causal relations by econometric models and cross-spectral methods[J]. Econometrica, 1969, **37**(3): 424-438.

[13] Schreiber T. Measuring information transfer[J]. Physical Review Letters, 2000, **85**(2): 461-464.

[14] Sugihara G, May R, Ye H, et al. Detecting causality in complex ecosystems[J]. Science, 2012, **338**(6106): 496-500.

[15] O′brien R M. A caution regarding rules of thumb for variance inflation factors[J]. Quality & Quantity, 2007, **41**(5): 673-690.

[16] Wang Y X, Wu H X, Dong J X, et al. Timexer: empowering transformers for time series forecasting with exogenous variables [J]. Advances in Neural Information Processing Systems, 2024, **37**: 469-498.

[17] Yuan X F, Qi S B, Wang Y L, et al. Quality variable prediction for nonlinear dynamic industrial processes based on temporal convolutional networks[J]. IEEE Sensors Journal, 2021, **21**(18): 20493-20503.

[18] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, **9**(8): 1735-1780.