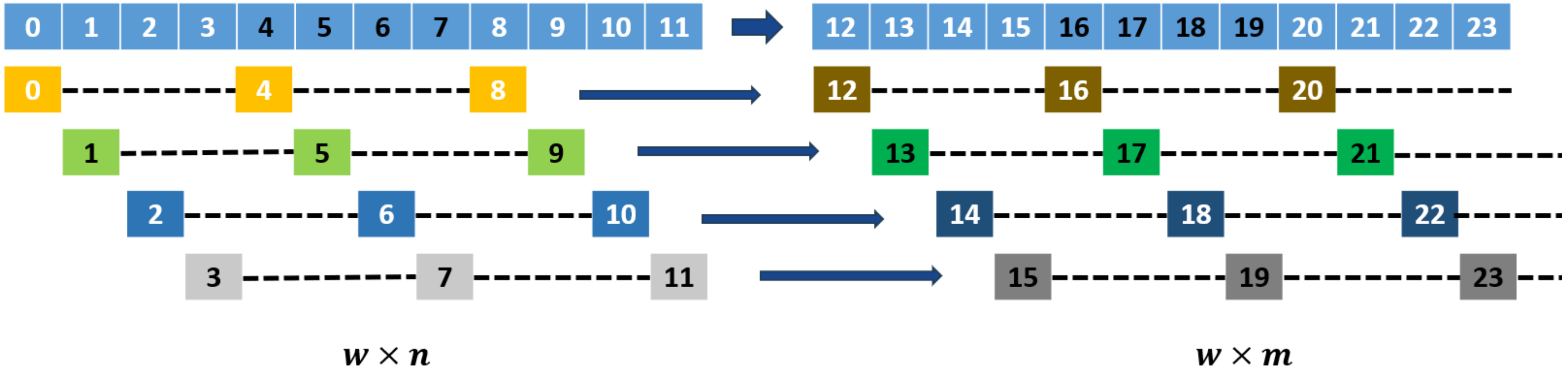
ICML2024 (Oral)   CCF-A

SparseTSF

Modeling Long-term **T**ime **S**eries **F**orecasting with 1k Parameters



ICML2024 (Oral)   CCF-A

SparseTSF

Modeling Long-term **T**ime **S**eries **F**orecasting with 1k Parameters

24.6.17

Presented by Yyyq

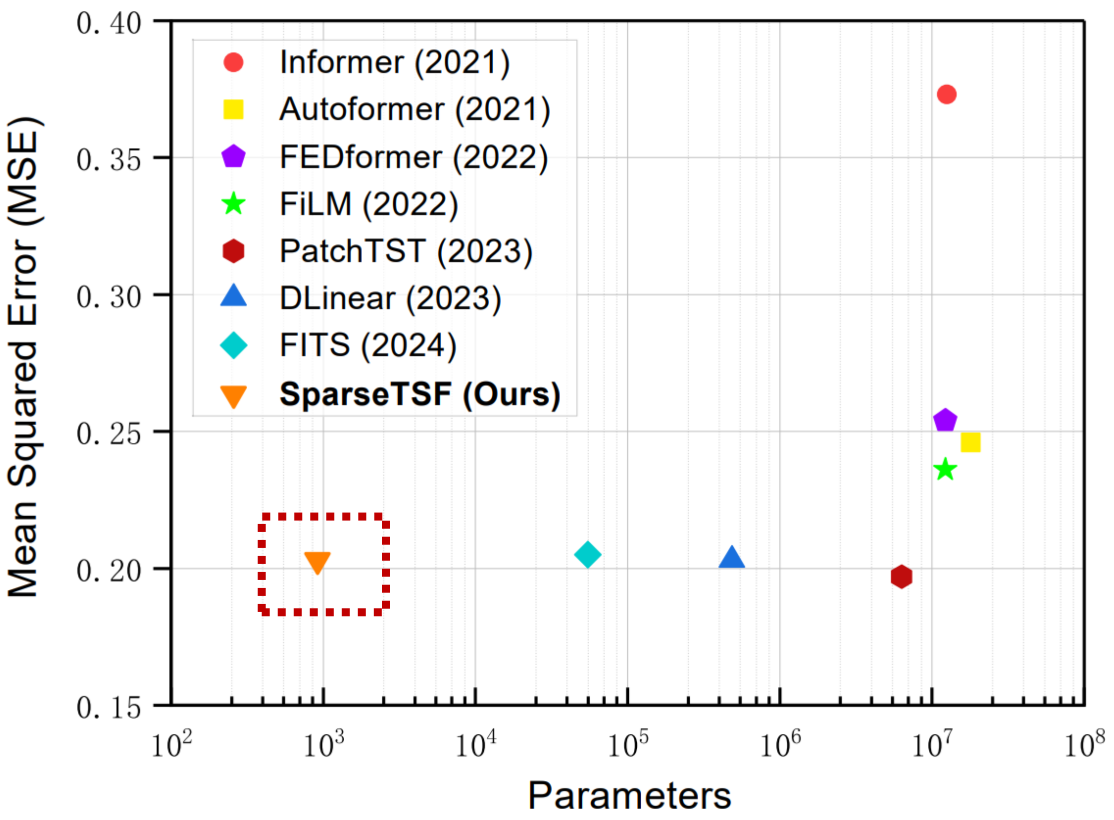
01

问题描述

- **LTSF（Long-term Time Series Forecasting）：**
  - 从更长的历史窗口中提取更广泛的时间依赖性
  - 数据固有的周期性和趋势性
- **Transformer-Based：**
  - 在计算资源受限的场景中可用性有限

02 创新点

- **稀疏技术**：跨周期稀疏预测技术
  - 解耦数据的周期性和趋势
  - 使用固定的周期性特征，专注预测趋势
- **SparseTSF 模型**：基于稀疏技术
  - 参数大小极端压缩
  - 具有竞争力的预测精度
  - 稳健的泛化能力



03 算法实现：预备知识

- **长时预测**
  - L 预测 H，C 个通道
- **通道独立策略**
  - 为每一个单变量序列找到一个共享函数（函数共享，参数独立）
  - 在单变量序列中建模长期依赖关系
- **数据要具有先验知识**
  - 表现出恒定的周期性，已知一个周期 w

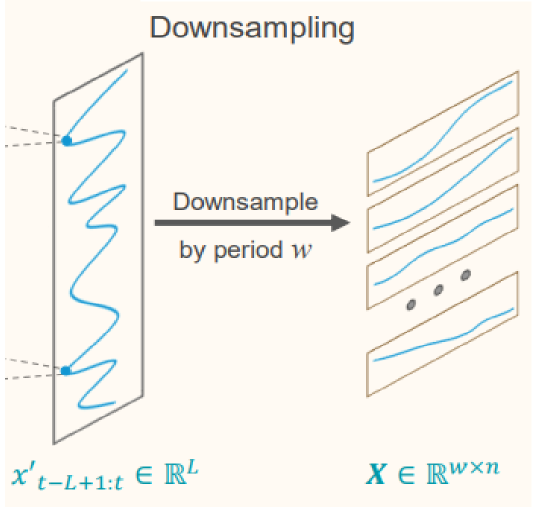
03 算法实现：稀疏技术 Cross-Period Sparse Forecasting

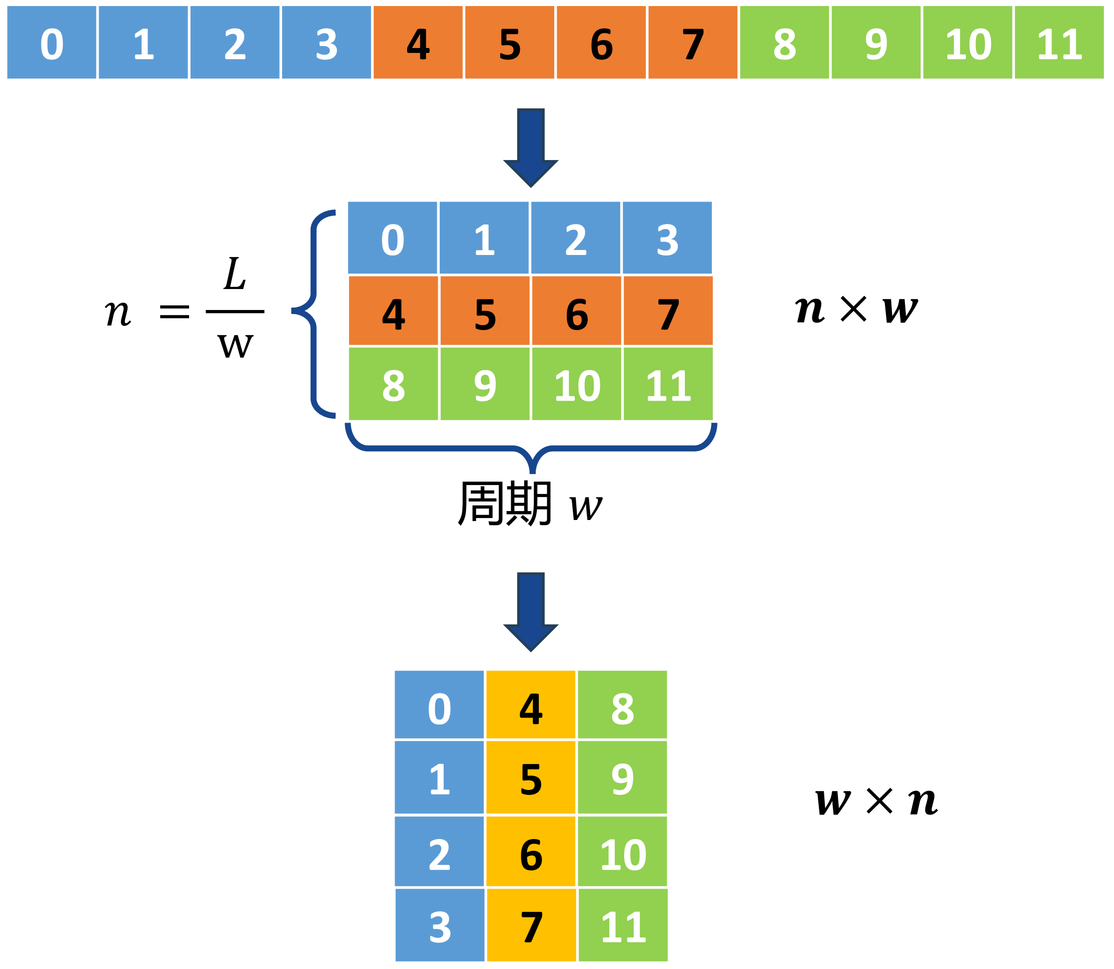
➤ **下采样**

- 单变量时序 L，已知周期 w

  $\rightarrow w$个子序列，每个序列长度 $n = \frac{L}{w}$

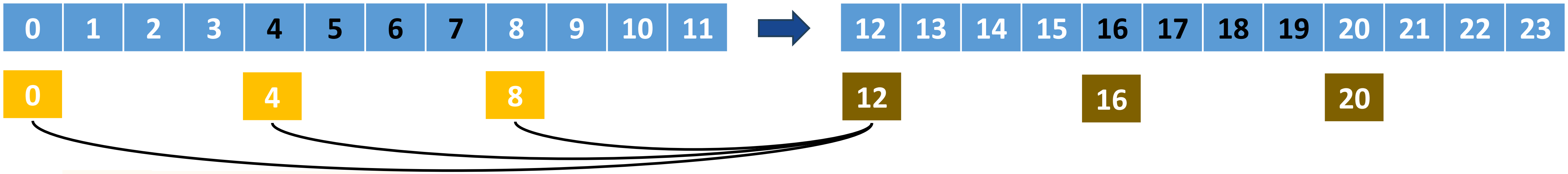
- 得到 $n \times w$ 的矩阵
- 转置为 $w \times n$

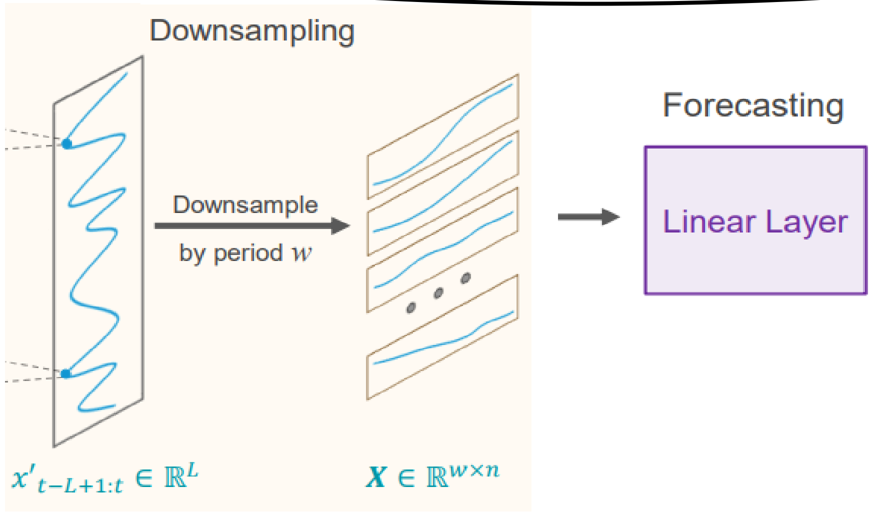




03 算法实现：稀疏技术 Cross-Period Sparse Forecasting

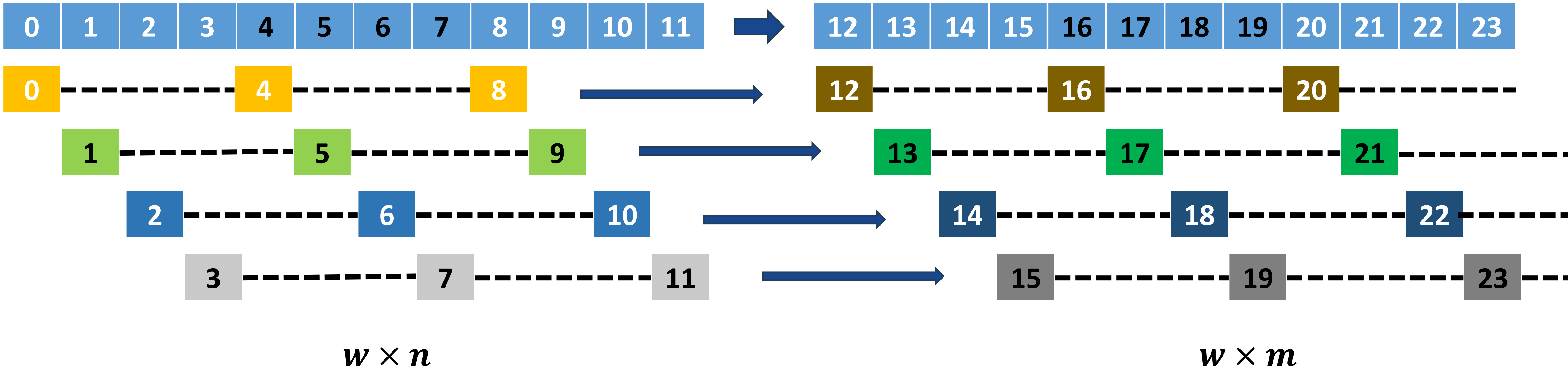
- **下采样：** $\mathrm{L} \rightarrow w \times n$
- **稀疏滑动预测：**
  - 应用线性层进行预测：$n \rightarrow m\ (m = \frac{H}{\mathrm{w}})$





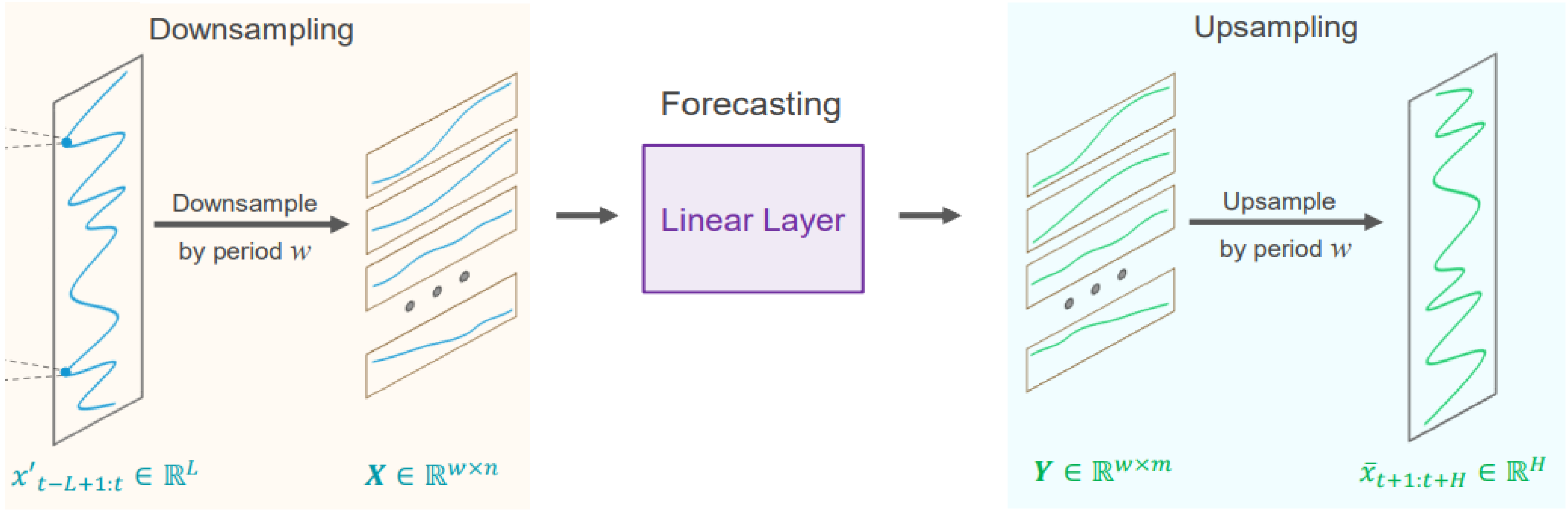
03 算法实现：稀疏技术 Cross-Period Sparse Forecasting

- 下采样： $L \rightarrow w \times n$
- 稀疏滑动预测： $w \times n \rightarrow w \times m$



03 算法实现：稀疏技术 Cross-Period Sparse Forecasting

- **下采样：** $\mathrm{L} \rightarrow w \times n$
- **稀疏滑动预测：** $w \times n \rightarrow w \times m$
- **上采样**
  - 将矩阵重塑为长度为H的预测序列： $w \times m \rightarrow H$

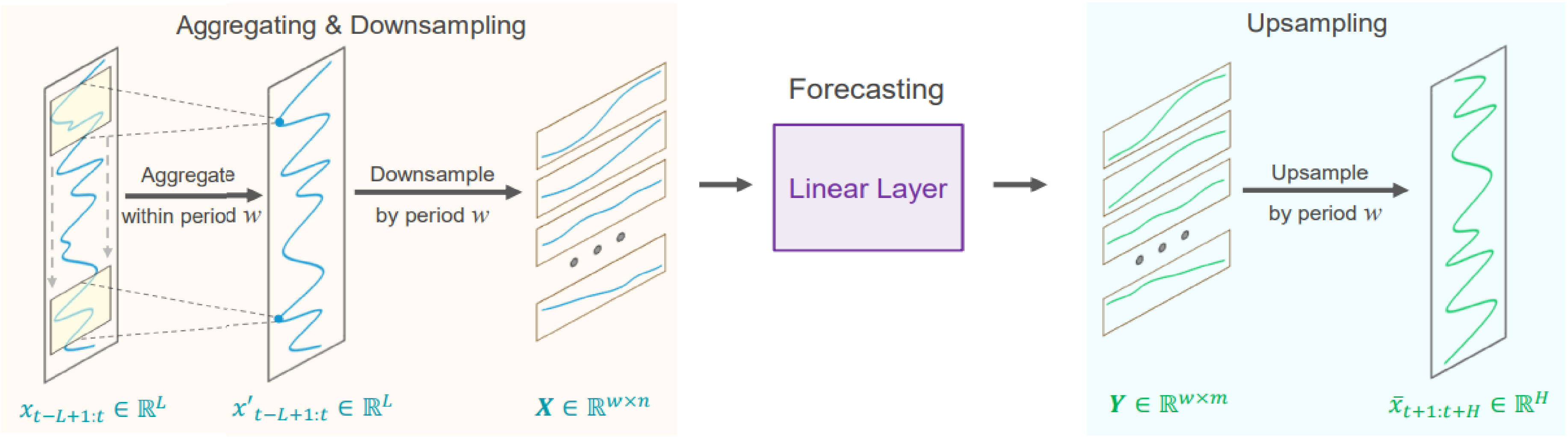


① **信息丢失**

② **放大异常值的影响**

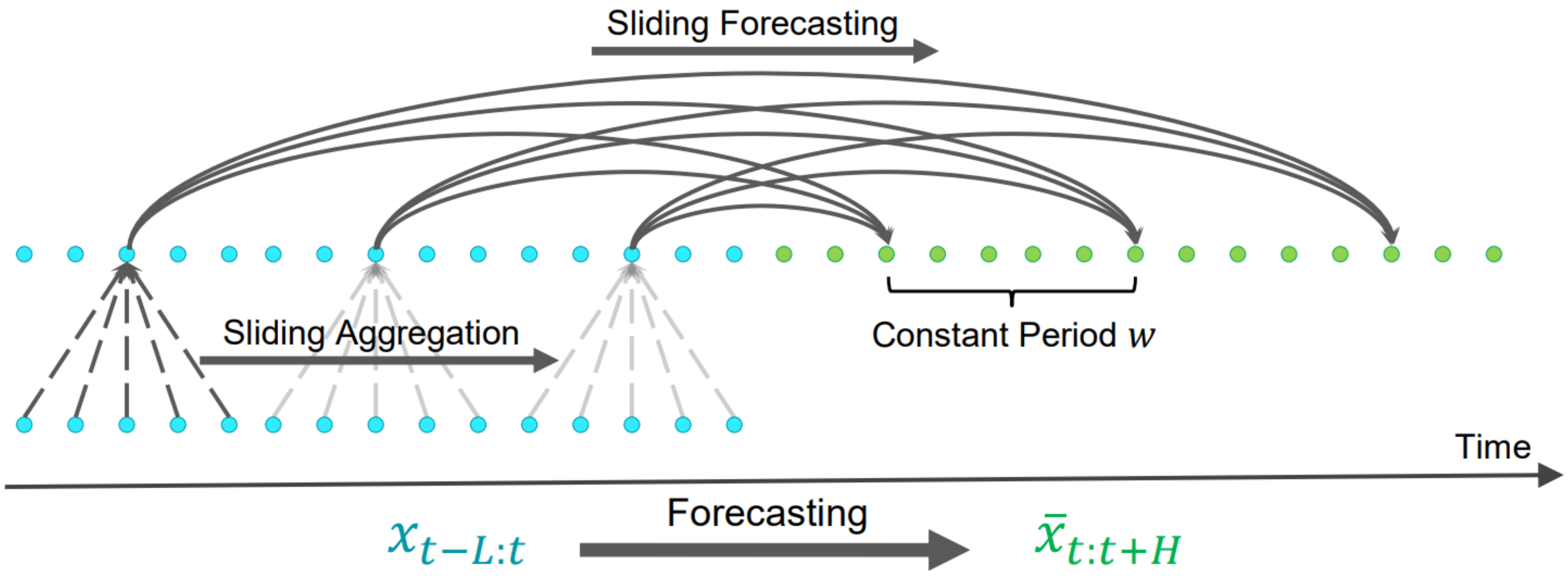
03 算法实现：SparseTSF

- **滑动聚合**
  - $x^{(i)}_{t-L+1:t} = x^{(i)}_{t-L+1:t} + \boxed{\text{Conv1D}(x^{(i)}_{t-L+1:t})}$ kernel size of $2 \times \lfloor \frac{w}{2} \rfloor + 1.$
- **下采样：** $\mathrm{L} \rightarrow w \times n$
- **稀疏滑动预测：** $w \times n \rightarrow w \times m$
- **上采样**
  - 将矩阵重塑为长度为H的预测序列： $w \times m \rightarrow H$



03 算法实现：SparseTSF

- 滑动聚合
  - $x^{(i)}_{t-L+1:t} = x^{(i)}_{t-L+1:t} + \boxed{\text{Conv1D}(x^{(i)}_{t-L+1:t})}$ kernel size of $2 \times \lfloor \frac{w}{2} \rfloor + 1.$
- 下采样： $L \rightarrow w \times n$
- 稀疏滑动预测：$w \times n \rightarrow w \times m$
- 上采样： $w \times m \rightarrow H$



04 理论分析：参数效率

- **总参数量**
  - $\left\lfloor \frac{L}{w} \right\rfloor \times \left\lfloor \frac{H}{w} \right\rfloor + 2 \times \left\lfloor \frac{w}{2} \right\rfloor + 1.$
- **线性层：$\boldsymbol{Y} = \boldsymbol{XW}$**
  - 权重矩阵 W ： $n \times m = \left\lfloor \frac{L}{\mathrm{w}} \right\rfloor \times \left\lfloor \frac{H}{\mathrm{w}} \right\rfloor$；
- **卷积核：$\boldsymbol{Y} = \boldsymbol{C}\mathrm{onv}\mathbf{1}\boldsymbol{D}(\boldsymbol{X}, \boldsymbol{K})$**
  - 卷积和大小 K： $K = 1 + 2 \times \left\lfloor \frac{\mathrm{w}}{2} \right\rfloor$

具体示例

假设 $L = 720$，$H = 720$，$w = 24$：

1. **线性层的参数数量**：

$$\left\lfloor \frac{720}{24} \right\rfloor \times \left\lfloor \frac{720}{24} \right\rfloor = 30 \times 30 = 900$$

2. **卷积层的参数数量**：

$$1 + 2 \times \left\lfloor \frac{24}{2} \right\rfloor = 1 + 2 \times 12 = 25$$

3. **总参数数量**：

$$900 + 25 = 925$$

这样，通过公式我们可以看到 SparseTSF 模型的参数总数远小于传统模型。

04 理论分析：稀疏策略的有效性

- **基于恒定的周期性：**

L预测H

$$x_{t+1:t+H} = f(x_{t-L+1:t})$$

➡

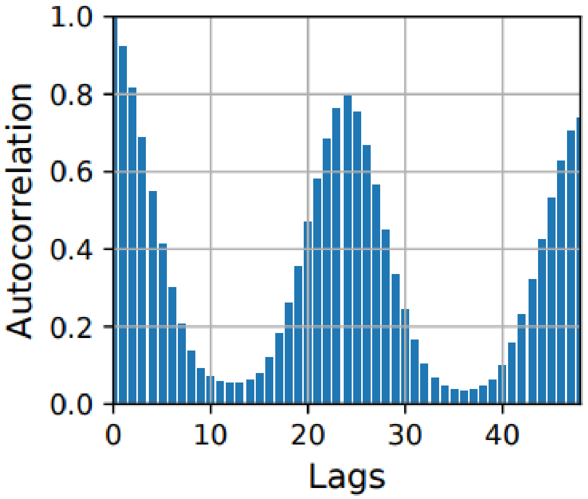
滑动预测：n预测m

$$x'_{t+1:t+m} = f(x'_{t-n+1:t})$$
$$p'_{t+1:t+m} + t'_{t+1:t+m} = f(p'_{t-n+1:t} + t'_{t-n+1:t})$$
$$p'_i = p'_j$$

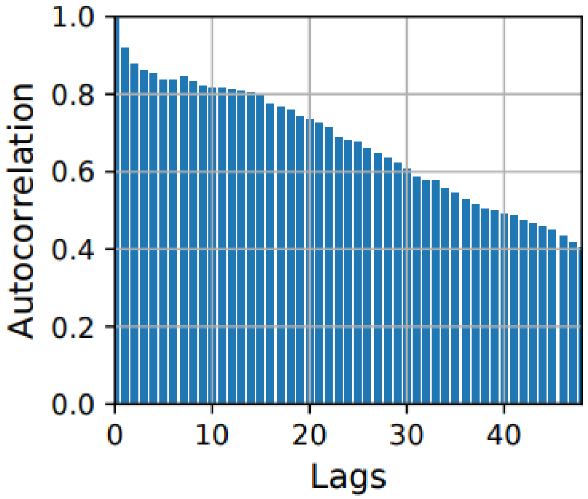
- 模型的输出是未来的周期性和趋势成分之和
- 不显式地对周期性成分进行建模，而是将其作为一个已知的参考
- 通过分离周期性成分，模型更专注于趋势变化

04 理论分析：稀疏策略的有效性

- **基于恒定的周期性：** $p'_i = p'_j$
- **自相关函数ACF**
  - 当滞后时间 $k = w$ 时，ACF 通常会显示显著的峰值：**周期性特征**
  - 下采样序列的自相关图中，没有显示出显著的周期性峰值：**保留了趋势特征**



(a) Original



(b) Downsampled

05

实验

➢ **数据集：**小时级，具有日周期性

Datasets	ETTh1 & ETTh2	Electricity	Traffic
Channels	7	321	862
Frequency	hourly	hourly	hourly
Timesteps	17,420	26,304	17,544

➢ **主要基线：**

- FITS（ICLR2024）：FITS_Modeling Time Series with 10k Parameters
- patchTST，DLinear

05 实验1：预测效果对比

Dataset	ETTh1				ETTh2				Electricity				Traffic			
Horizon	96	192	336	720	96	192	336	720	96	192	336	720	96	192	336	720
Informer (2021)	0.865	1.008	1.107	1.181	3.755	5.602	4.721	3.647	0.274	0.296	0.300	0.373	0.719	0.696	0.777	0.864
Autoformer (2021)	0.449	0.500	0.521	0.514	0.358	0.456	0.482	0.515	0.201	0.222	0.231	0.254	0.613	0.616	0.622	0.660
Pyraformer (2022b)	0.664	0.790	0.891	0.963	0.645	0.788	0.907	0.963	0.386	0.386	0.378	0.376	2.085	0.867	0.869	0.881
FEDformer (2022b)	0.376	0.420	0.459	0.506	0.346	0.429	0.496	0.463	0.193	0.201	0.214	0.246	0.587	0.604	0.621	0.626
FiLM (2022a)	0.371	0.414	0.442	0.465	0.284	0.357	0.377	0.439	0.154	0.164	0.188	0.236	0.416	0.408	0.425	0.520
TimesNet (2023)	0.384	0.436	0.491	0.521	0.340	0.402	0.452	0.462	0.168	0.184	0.198	0.220	0.593	0.617	0.629	0.640
PatchTST (2023)	**0.370**	0.413	**0.422**	0.447	**0.274**	0.341	**0.329**	0.379	**0.129**	**0.147**	**0.163**	**0.197**	**0.360**	**0.379**	**0.392**	**0.432**
DLinear (2023)	0.374	**0.405**	0.429	0.440	0.338	0.381	0.400	0.436	0.140	0.153	0.169	**0.203**	0.410	0.423	0.435	0.464
FITS (2024)	0.375	0.408	0.429	**0.427**	**0.274**	**0.333**	0.340	**0.374**	**0.138**	0.152	0.166	0.205	0.401	0.407	0.420	0.456
SparseTSF (ours)	**0.359** ±0.006	**0.397** ±0.002	**0.404** ±0.001	**0.417** ±0.001	**0.267** ±0.005	**0.314** ±0.003	**0.312** ±0.004	**0.370** ±0.001	**0.138** ±0.001	**0.146** ±0.001	**0.164** ±0.001	**0.203** ±0.001	**0.382** ±0.001	**0.388** ±0.001	**0.402** ±0.001	**0.445** ±0.002
Imp.	+0.011	+0.008	+0.018	+0.010	+0.007	+0.019	+0.017	+0.004	-0.009	+0.001	-0.001	-0.006	-0.022	-0.009	-0.010	-0.013

- **SparseTSF在所有场景中都排在前两名之内**
- **SparseTSF结果的标准偏差非常小**

05

实验2：效率优势

Model	Parameters	MACs	Max Mem.(MB)	Epoch Time(s)
Informer (2021)	12.53 M	3.97 G	969.7	70.1
Autoformer (2021)	12.22 M	4.41 G	2631.2	107.7
FEDformer (2022b)	17.98 M	4.41 G	1102.5	238.7
FiLM (2022a)	12.22 M	4.41 G	1773.9	78.3
PatchTST (2023)	6.31 M	11.21 G	10882.3	290.3
DLinear (2023)	485.3 K	156.0 M	123.8	25.4
FITS (2024)	10.5 K	79.9 M	496.7	35.0
SparseTSF (Ours)	**0.92 K**	**12.71 M**	125.2	31.3

- 静态指标：参数总量 和 MACs乘法累加操作总量 **——小10倍以上**
- 运行指标：最大内存和Epoch Time **——与现有的轻量级模型相媲美**

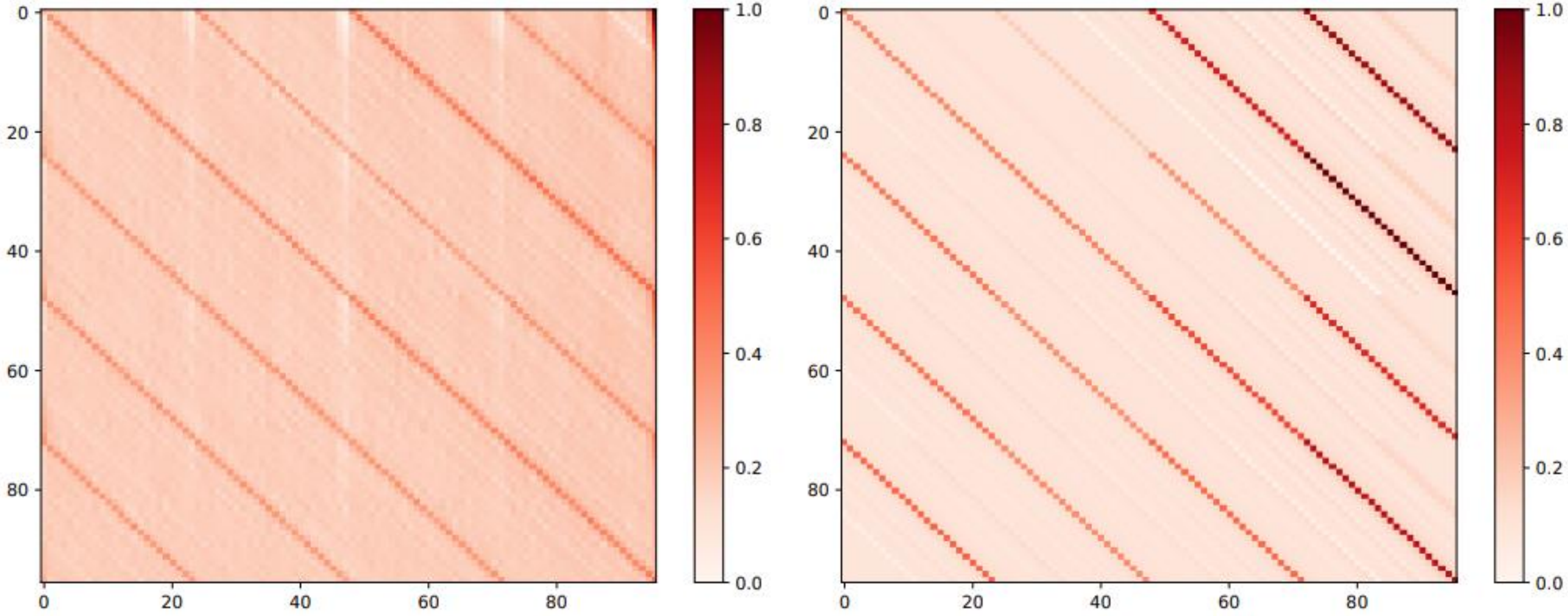
实验3：稀疏技术分析——有效性

Dataset	ETTh1				ETTh2			
Horizon	96	192	336	720	96	192	336	720
Linear	0.371	0.460	0.417	0.424	**0.257**	0.337	0.336	0.391
+sparse	**0.359**	**0.397**	**0.404**	**0.417**	0.267	**0.314**	**0.312**	**0.370**
Boost	3.3%	13.8%	3.1%	1.7%	-3.9%	6.9%	7.1%	5.3%
Transformer	0.697	0.732	0.714	0.770	0.340	**0.376**	0.366	0.468
+sparse	**0.406**	**0.442**	**0.446**	**0.489**	**0.322**	0.380	**0.353**	**0.432**
Boost	41.7%	39.6%	37.5%	36.5%	5.2%	-1.0%	3.6%	7.7%
GRU	0.415	0.529	0.512	0.620	0.296	0.345	0.363	0.454
+sparse	**0.356**	**0.391**	**0.437**	**0.455**	**0.282**	**0.332**	**0.356**	**0.421**
Boost	14.1%	26.1%	14.7%	26.7%	4.8%	3.7%	1.9%	7.2%

- SparseTSF的核心：稀疏技术 + 单层线性模型
- **线性模型平均提高了4.7%，Transformer提高了21.4%，GRU提高了12.4%**

05

实验3：稀疏技术分析——权重矩阵可视化



(a) Linear　　(b) SparseTSF

05 实验4：超参数w的影响

Horizon	SparseTSF (w=6)	SparseTSF (w=12)	SparseTSF (w=24)	SparseTSF (w=48)	FITS (2024)	DLinear (2023)	PatchTST (2023)
96	0.376	0.369	**0.359**	0.380	0.375	0.374	0.370
192	0.410	0.402	**0.397**	0.400	0.408	0.405	0.413
336	0.408	0.406	0.404	**0.399**	0.429	0.429	0.422
720	0.427	0.423	**0.417**	0.427	0.427	0.440	0.447
Avg.	0.405	0.400	**0.394**	0.402	0.410	0.412	0.413

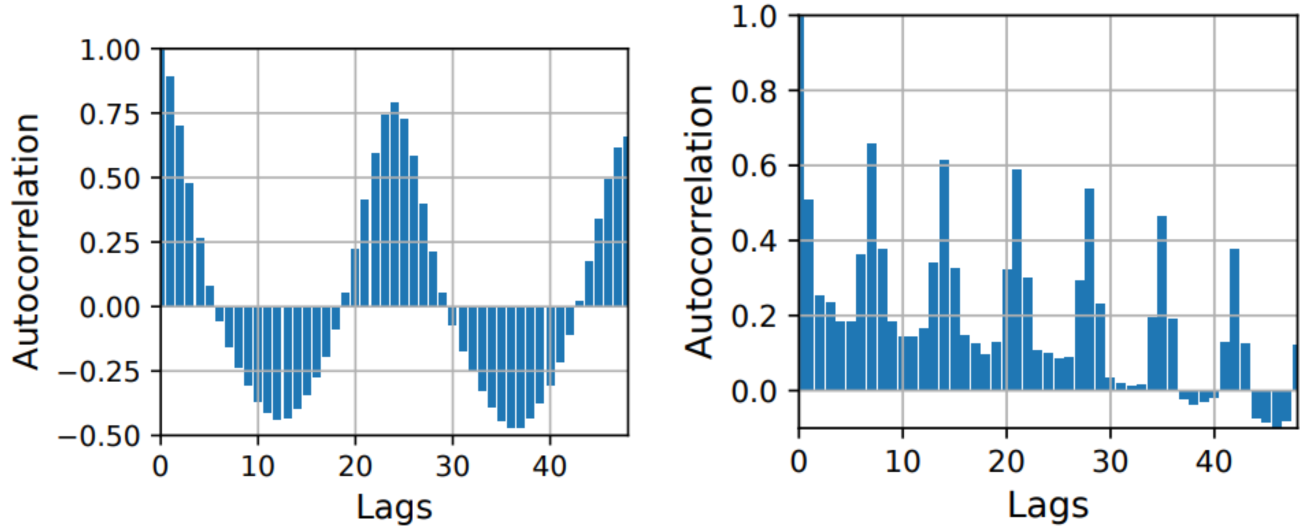
05

实验5：在不同数据集上的泛化能力

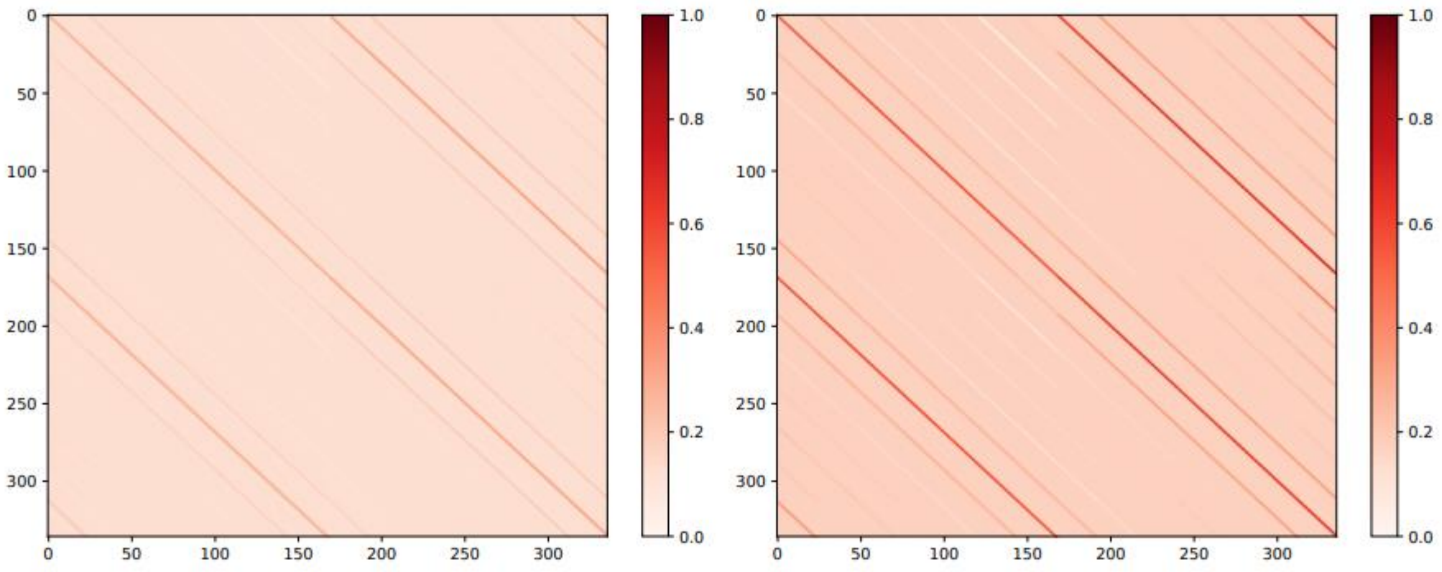
Dataset	ETTh2 → ETTh1				Electricity → ETTh1			
Horizon	96	192	336	720	96	192	336	720
Informer (2021)	0.844	0.921	0.898	0.829	\	\	\	\
Autoformer (2021)	0.978	1.058	0.944	0.921	\	\	\	\
FEDformer (2022b)	0.878	0.927	0.939	0.967	\	\	\	\
FiLM (2022a)	0.876	0.904	0.919	0.925	\	\	\	\
PatchTST (2023)	0.449	0.478	0.482	0.476	0.400	0.424	0.475	0.472
DLinear (2023)	0.430	0.478	0.458	0.506	0.397	0.428	0.447	0.470
Fits (2024)	0.419	0.427	0.428	0.445	0.380	0.414	0.440	0.448
SparseTSF (Ours)	**0.370**	**0.401**	**0.412**	**0.419**	**0.373**	**0.409**	**0.433**	**0.439**

PatchTST (2023)	**0.370**	0.413	**0.422**	0.447
DLinear (2023)	0.374	**0.405**	0.429	0.440
FITS (2024)	0.375	0.408	0.429	**0.427**
SparseTSF (ours)	**0.359** ±0.006	**0.397** ±0.002	**0.404** ±0.001	**0.417** ±0.001

06 模型限制：多周期问题



(a) Original (b) Downsampled



(a) Linear (b) SparseTSF

06 模型限制：超长周期问题

Table 9: MSE results of SparseTSF on ultra-long period datasets with varied hyperparameters w. The forecast horizon is set as 720.

Dataset	Parameter w							
	144	72	48	24	12	6	2	1
ETTm1	0.450	0.450	0.422	0.422	0.421	**0.415**	**0.415**	0.429
ETTm2	0.375	0.371	0.373	0.352	0.354	**0.349**	**0.349**	0.357
Weather	0.332	0.329	0.325	0.321	0.319	0.319	**0.318**	0.322

- ETTm1、ETTm2和Weather数据集，主要周期分别长达96和144
- 在SparseTSF框架内采用更密集的连接
- 适度的稀疏连接（w=1时性能不如稀疏连接预测）可以提高模型的预测准确性

07 未来工作：不同回溯长度的影响

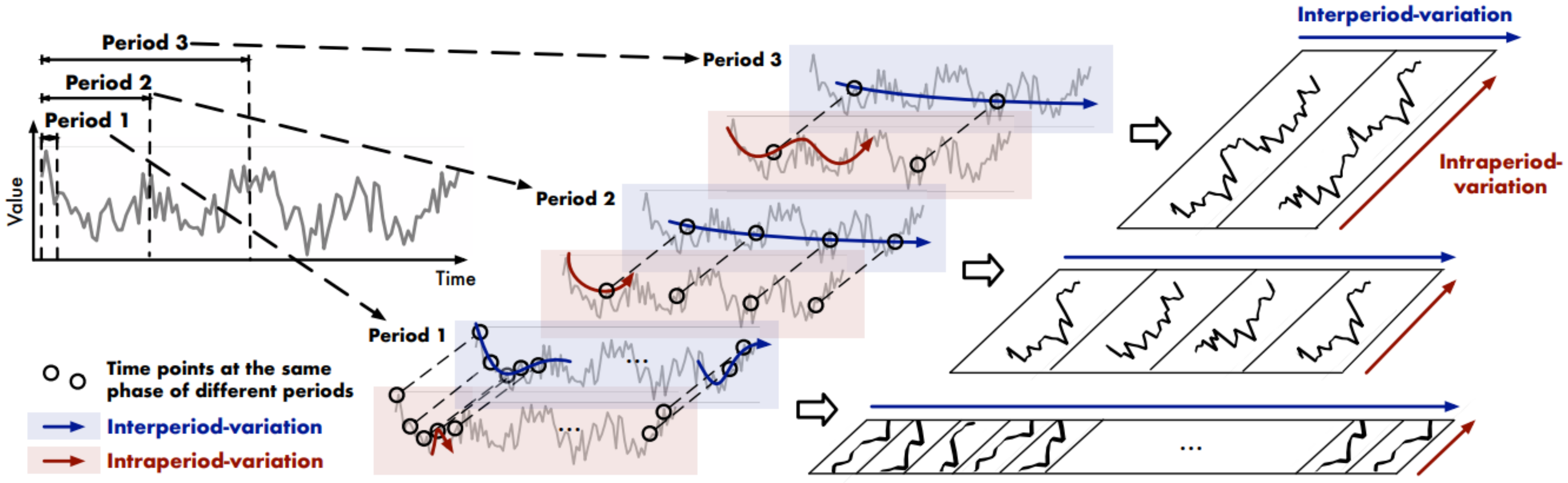
Dataset	ETTh1				ETTh2				Electricity				Traffic			
Horizon \ Look-back	96	192	336	720	96	192	336	720	96	192	336	720	96	192	336	720
96	0.380	0.371	0.393	0.354	0.288	0.285	0.272	0.278	0.209	0.160	0.146	0.138	0.672	0.455	0.412	0.383
192	0.433	0.434	0.418	0.398	0.363	0.346	0.323	0.315	0.202	0.166	0.154	0.147	0.608	0.453	0.415	0.388
336	0.447	0.420	0.390	0.405	0.366	0.335	0.314	0.311	0.217	0.184	0.172	0.164	0.609	0.468	0.428	0.403
720	0.451	0.426	0.413	0.418	0.407	0.389	0.372	0.371	0.259	0.223	0.210	0.205	0.650	0.493	0.462	0.446
Avg.	0.428	0.413	0.404	**0.394**	0.356	0.339	0.320	**0.319**	0.222	0.183	0.171	**0.163**	0.635	0.467	0.429	**0.405**

- ETTh1 (w = 24)：96的回溯长度可以获得很好的结果，因为它们完全包含每日周期模式。
- Traffic (w = 24，168)：96的回看不能覆盖整个周周期模式，从而导致性能显著下降。

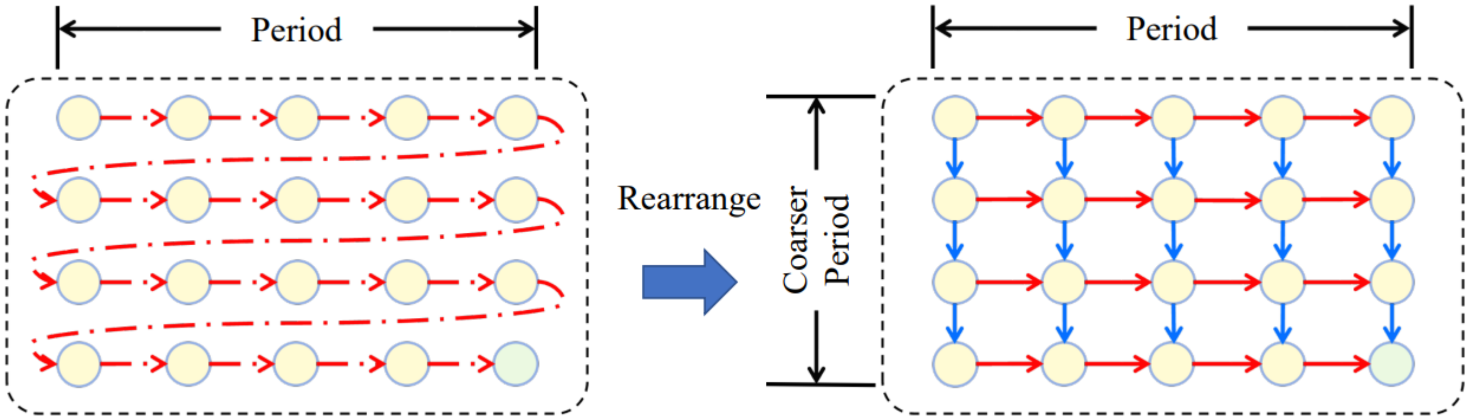
谢谢观看

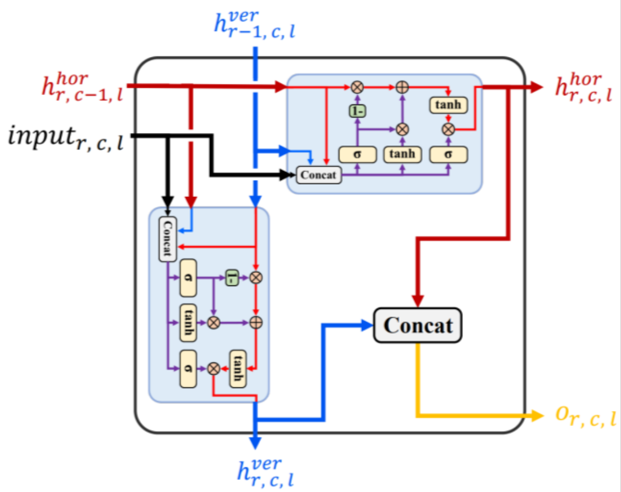
MANY THANKS !

24.6.18



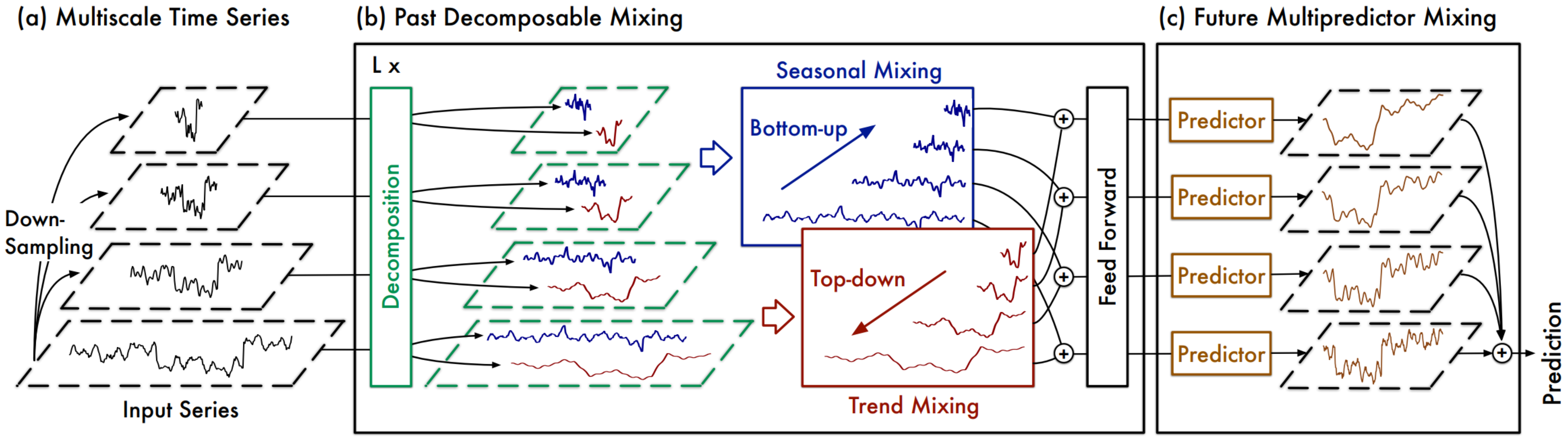
TimesNet





(b) HVGSU structure

WITRAN



TimeMixer

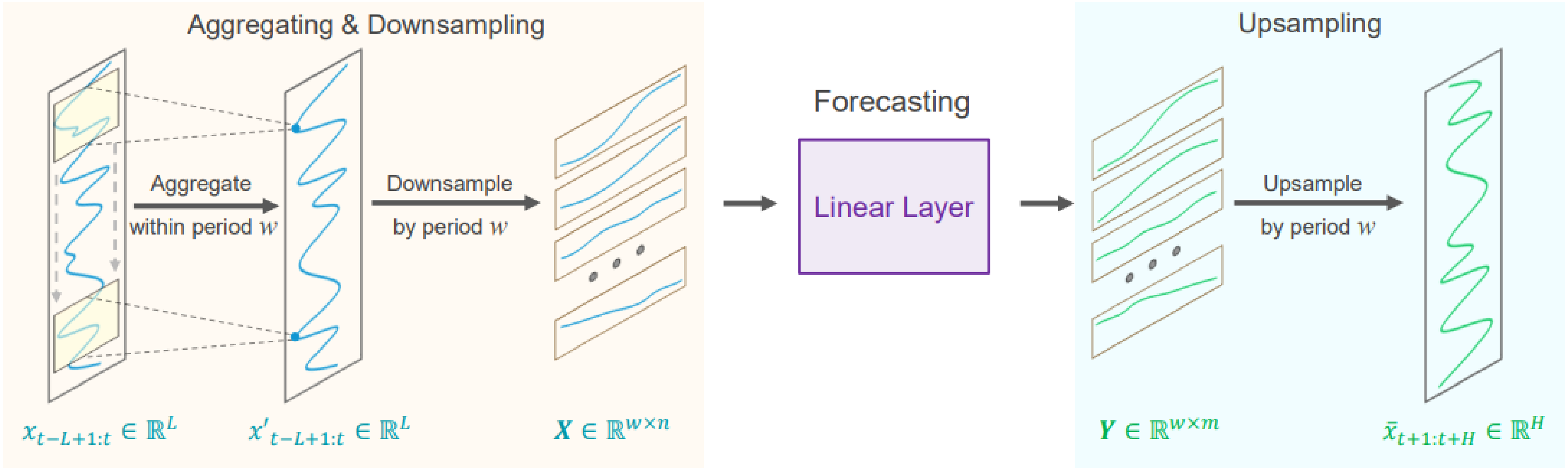


Figure 2: SparseTSF architecture.