# Fast and Slow Streams for Online Time Series Forecasting Without Information Leakage

Ying-Yee Ava Lau
yyalau@connect.ust.hk

Zhiwen Shao
zhiwen@ust.hk

Dit-Yan Yeung
dyyeung@cse.ust.hk

Department of Computer Science and Engineering

The Hong Kong University of Science and Technology

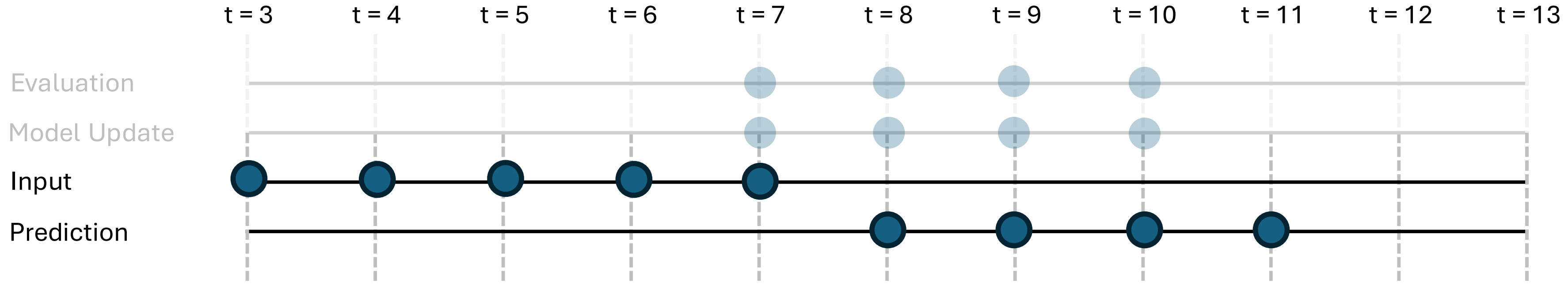
**Acknowledgement**
This work has been made possible by a Research Impact Fund project (R6003-21) and an Innovation and Technology Fund project (ITS/004/21FP) funded by the Hong Kong Government.

# Existing Online Time Series Forecasting (OTSF) Setting

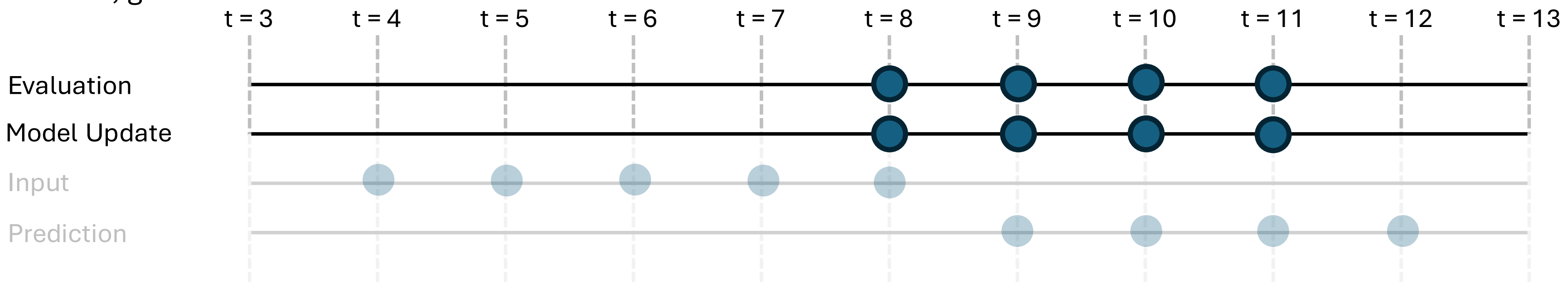
Look-back (Input) length $L = 5$

Horizon (Prediction) Length $H = 4$

**At t = 10,**



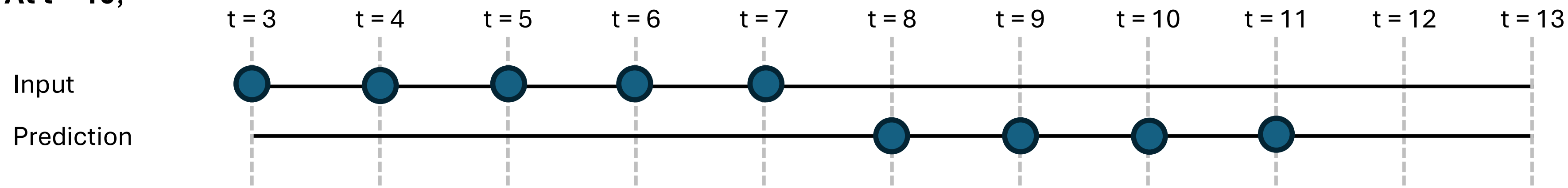

**At t = 11**, ground truth at t = 11 becomes available.



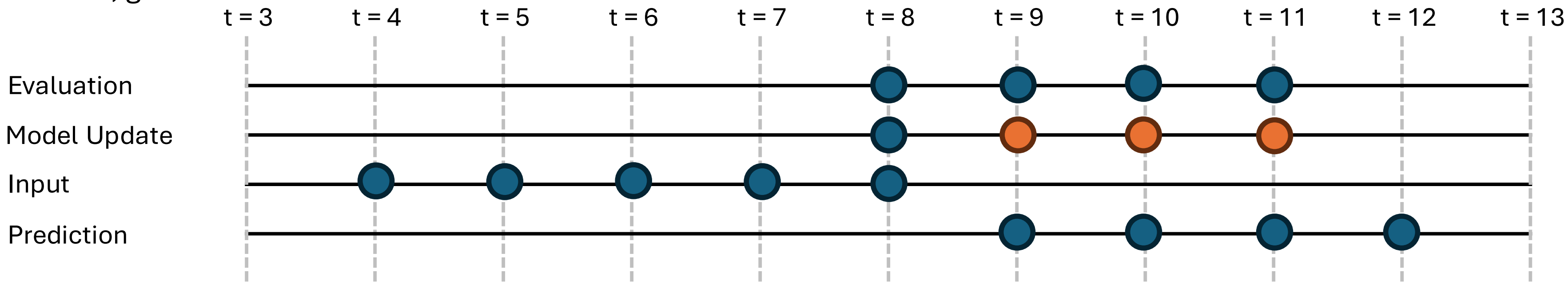
# Two Problems

1. Information Leakage
2. Only allow to forecast one future step

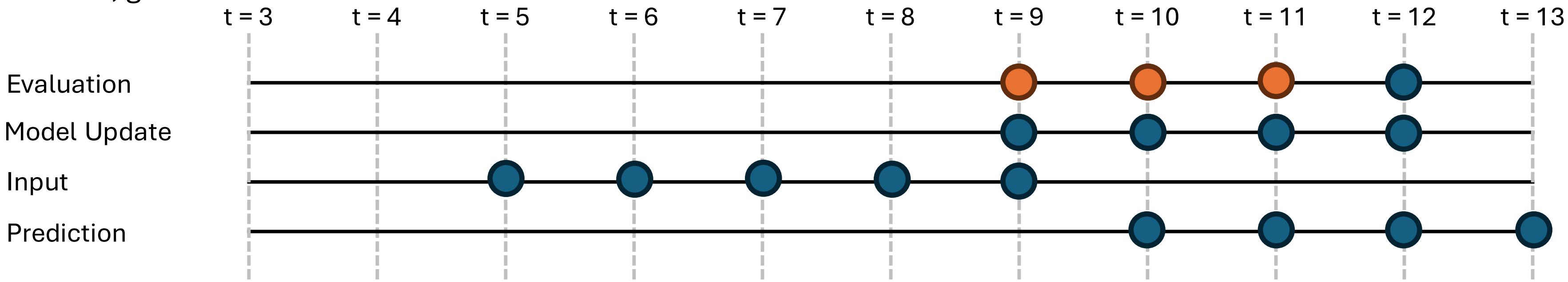
**At t = 10,**




**At t = 11**, ground truth at t = 11 becomes available.

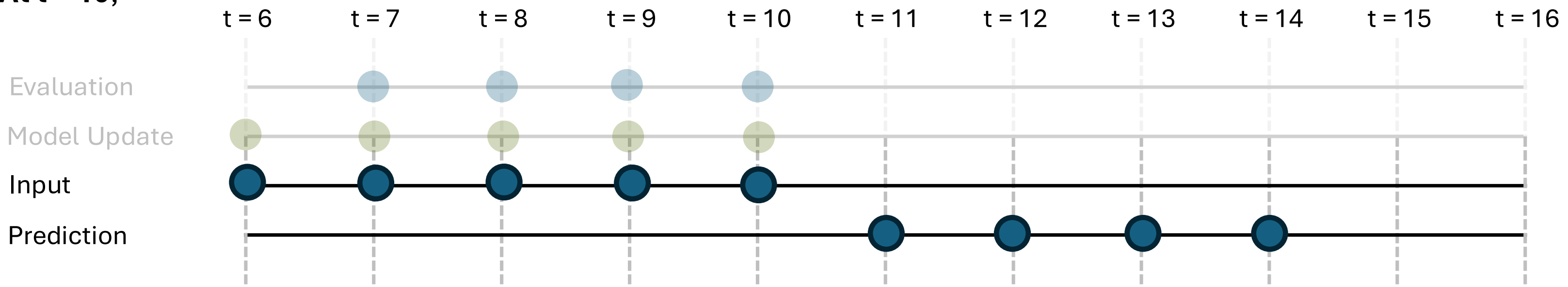



**At t = 12**, ground truth at t = 12 becomes available. Information leak occurs.

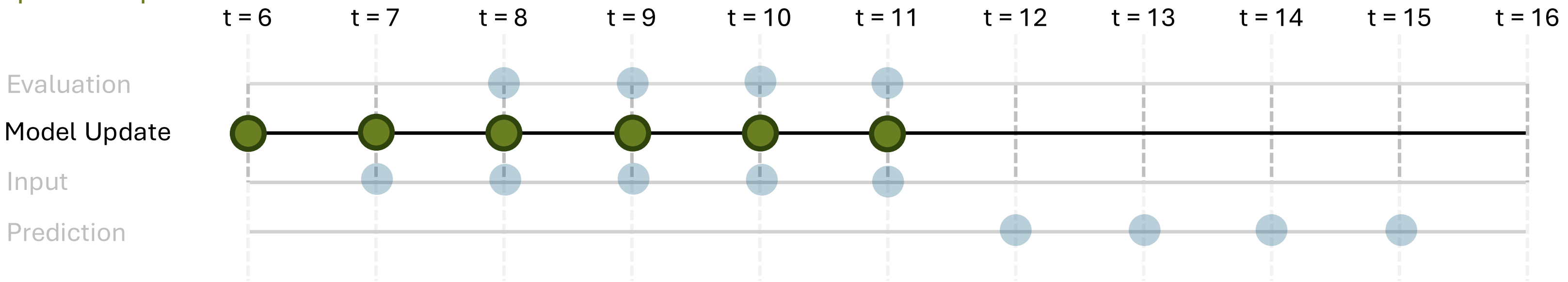


# Redefined OTSF Setting

**At t = 10,**

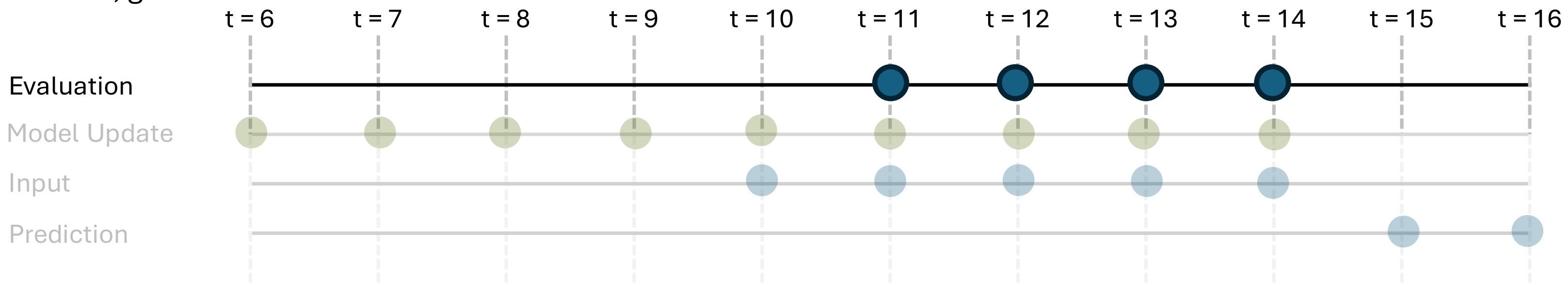



**At t = 11**, ground truth at t = 11 becomes available. The model can use any data points with timestamps t ≤ 11 to update its parameters.




……

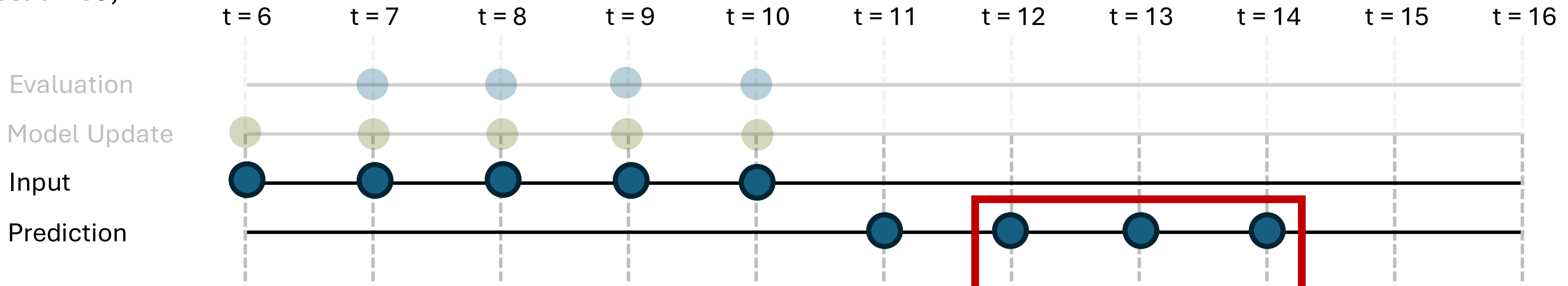
**At t = 14**, ground truth at t ≤ 14 are available.



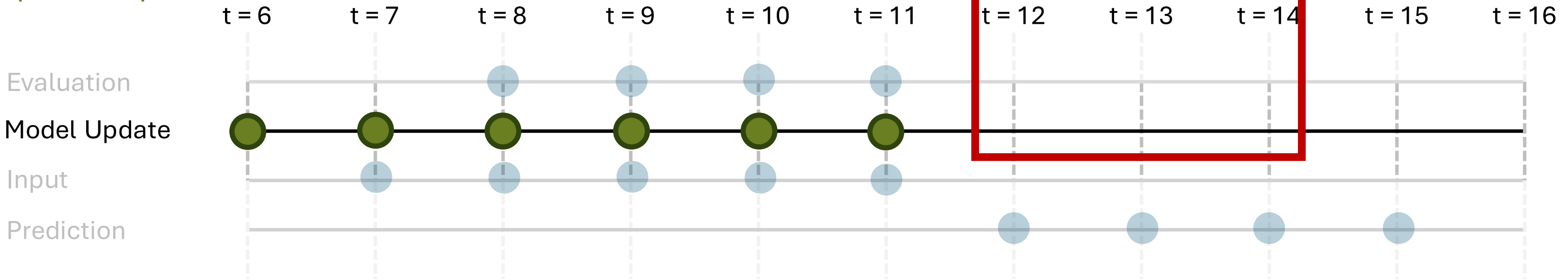
# Challenges

When $H > 1$, the ground truth for the most recent prediction is not fully accessible at the subsequent time step

**At t = 10,**

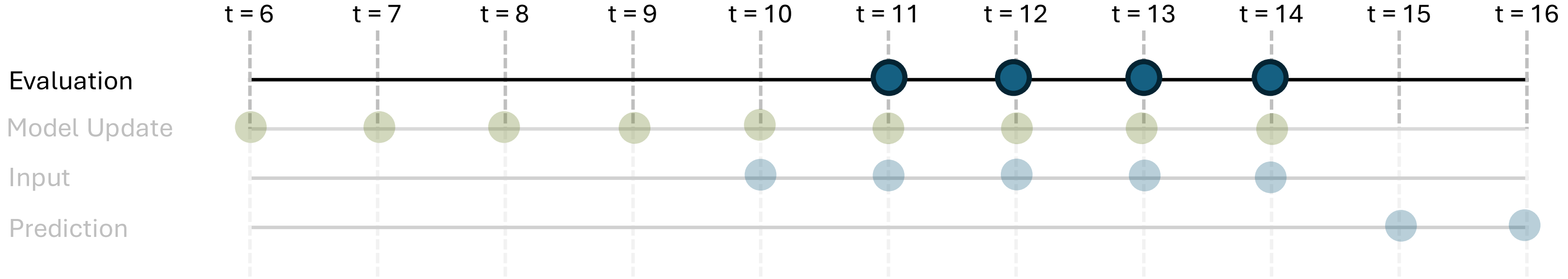



**At t = 11**, ground truth at t = 11 becomes available. The model can use any data points with timestamps t ≤ 11 to update its parameters.



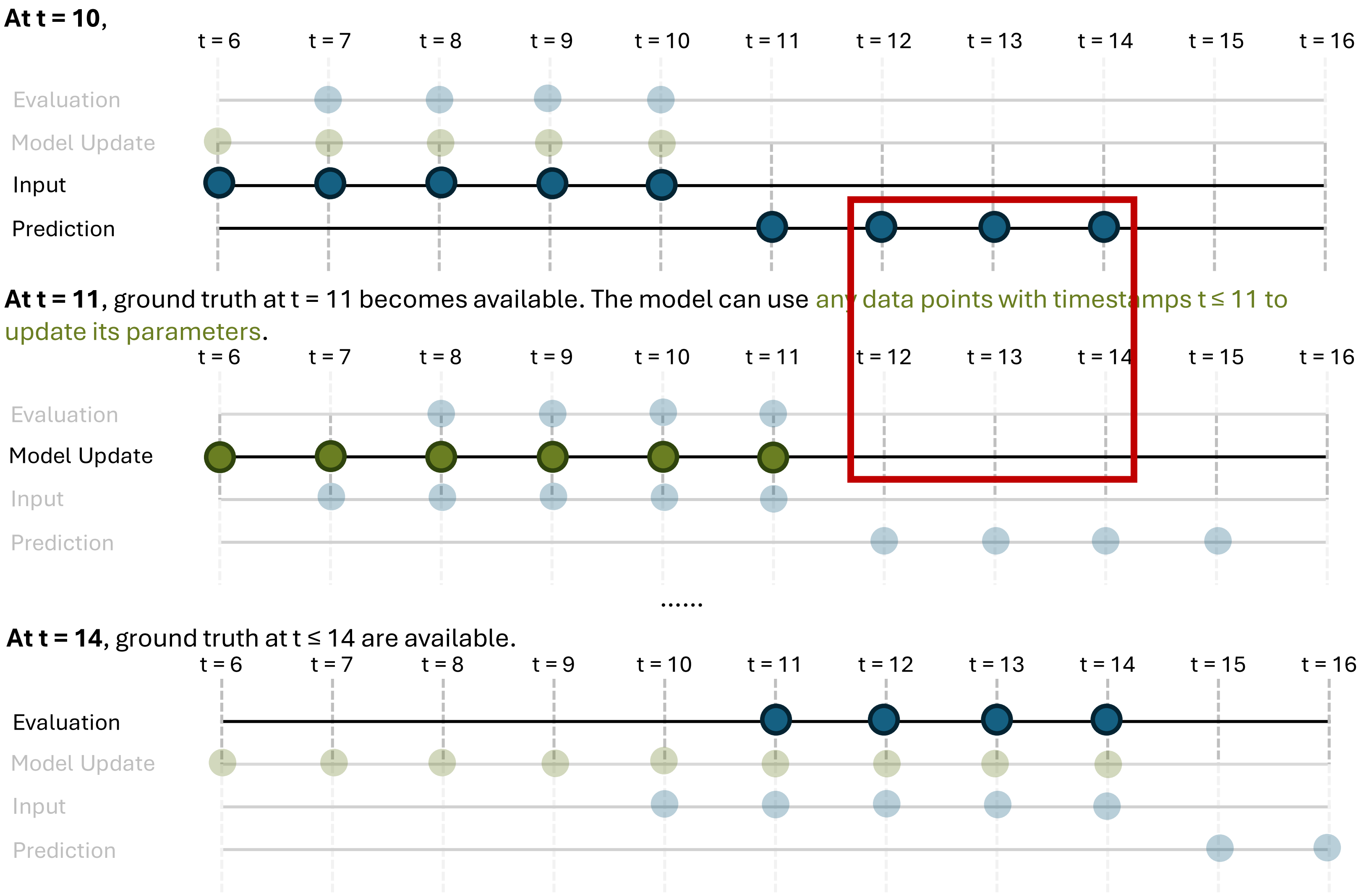

......

**At t = 14**, ground truth at t ≤ 14 are available.



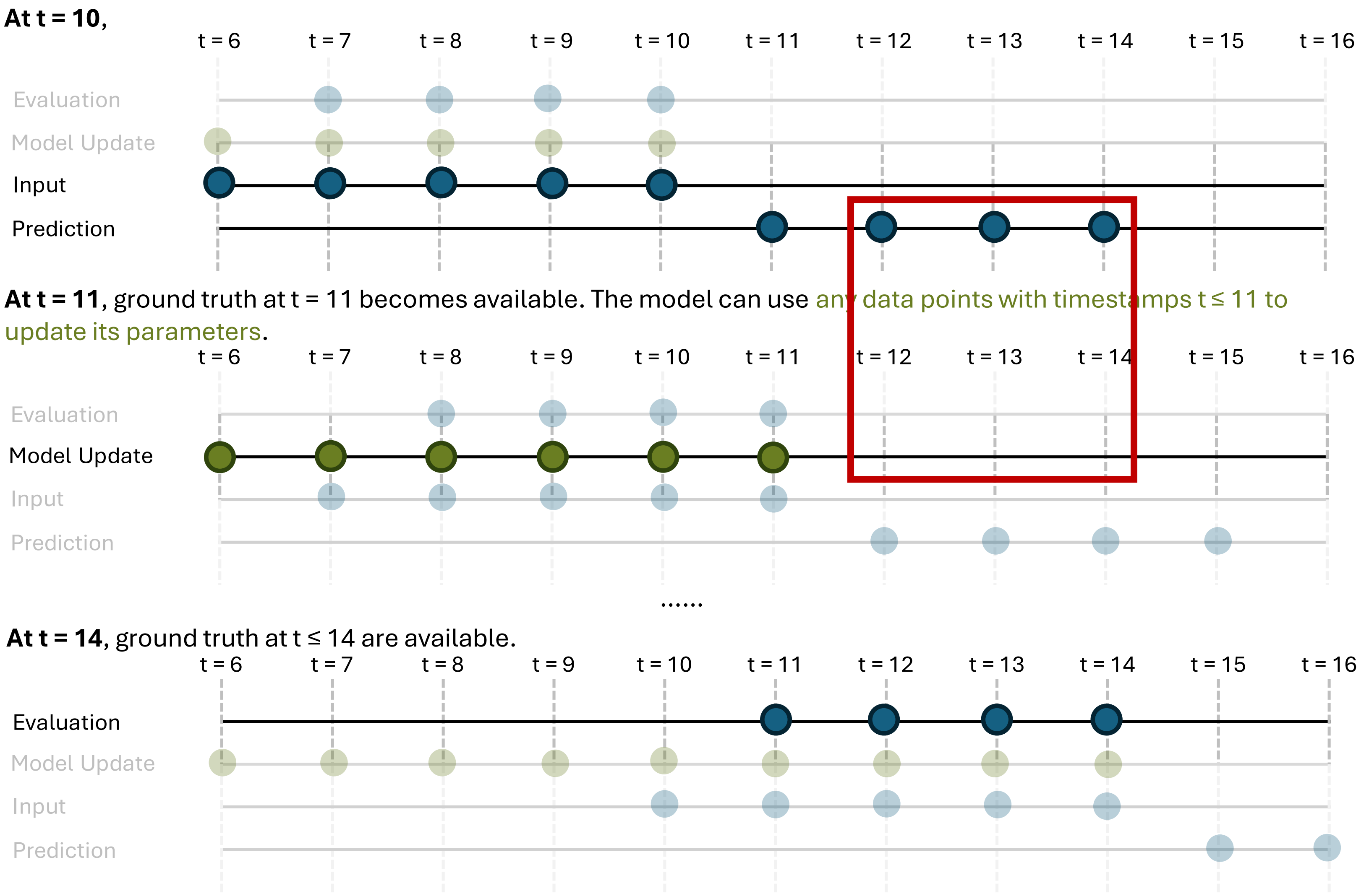
If we wait for all data to arrive…

The model cannot learn the latest information.



If we update with partial ground truth…

The model may capture noise or anomalies



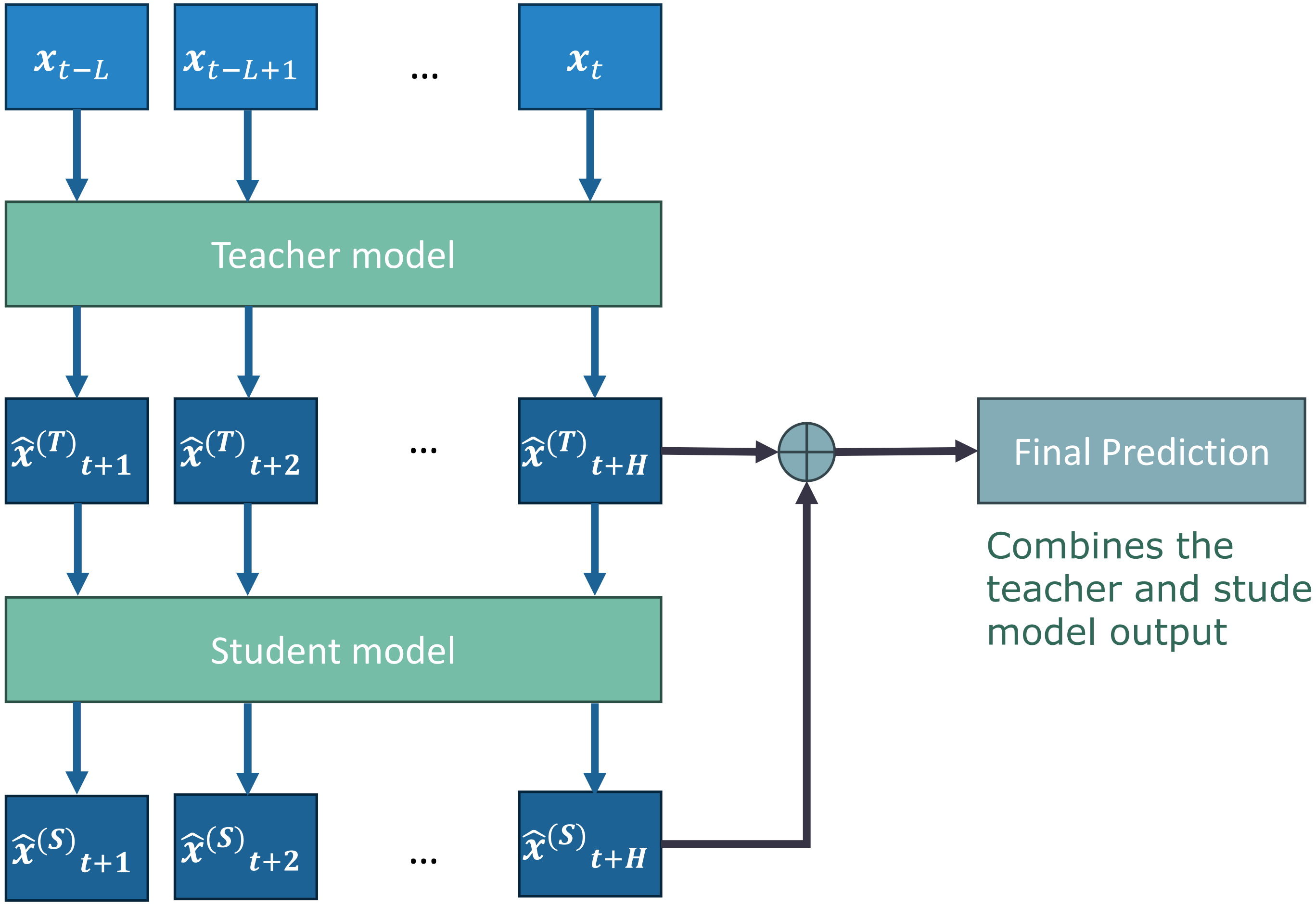
# Dual-Stream Online Forecasting (DSOF)

## Residual Learning Strategy

Input sequence

Generates coarse predictions

Refines predictions by estimating error




Combines the teacher and student model output

# Two Data Streams

**Slow stream**: Works with complete ground truth sequences

**Fast stream**: Enhance near-future forecasts without waiting for complete ground truth.

- Slow Data Stream with Experience Replay (ER)
  - Works with a First-In-First-Out buffer: When ground truth $\boldsymbol{x}_i$ arrives, append the newest sequence, and remove the oldest sample if buffer exceeds $N_{\boldsymbol{b}}$
  - Updates both the teacher and student model

**procedure** EXPERIENCE_REPLAY_UPDATE($\mathcal{B}, \boldsymbol{\theta}^{(T)}, \boldsymbol{\theta}^{(S)}$)

$(\mathcal{X}_L, \mathcal{X}_H) \sim \mathcal{B}$ ▷ *Sample a mini-batch from the buffer.*

$\hat{\mathcal{X}}_H[i] = f^{(i)}(\mathcal{X}_L; \boldsymbol{\theta}^{(T)}, \boldsymbol{\theta}^{(S)})$ (Equation 4) ▷ *Teacher-Student model makes a prediction.*

$\ell = \mathrm{MSE}(\hat{\mathcal{X}}_H, \mathcal{X}_H)$ ▷ *Compute the loss.*

$\boldsymbol{\theta}^{(T)} \leftarrow \boldsymbol{\theta}^{(T)} - \alpha_T \nabla_{\boldsymbol{\theta}^{(T)}} \ell(\boldsymbol{\theta}^{(T)}, \boldsymbol{\theta}^{(S)})$ ▷ *Update Teacher model's parameters.*

$\boldsymbol{\theta}^{(S)} \leftarrow \boldsymbol{\theta}^{(S)} - \alpha_S \nabla_{\boldsymbol{\theta}^{(S)}} \ell(\boldsymbol{\theta}^{(T)}, \boldsymbol{\theta}^{(S)})$ ▷ *Update Student model's parameters.*

**return** $\boldsymbol{\theta}^{(T)}, \boldsymbol{\theta}^{(S)}$

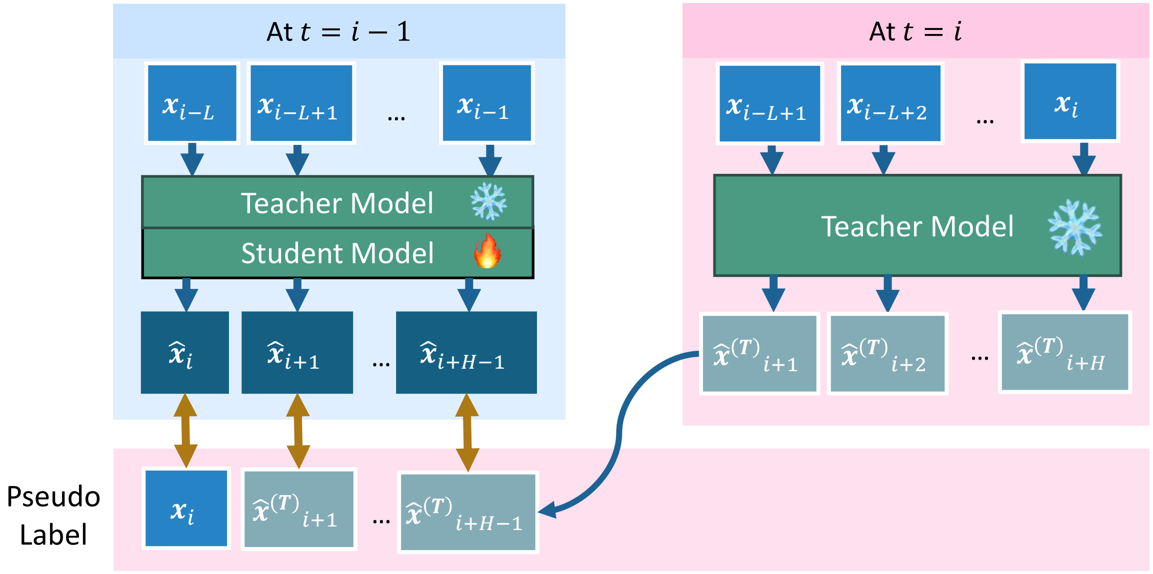
**end procedure**

# Two Data Streams

**Slow stream**: Works with complete ground truth sequences

**Fast stream**: Enhance near-future forecasts without waiting for complete ground truth.

- Fast Data Stream with Temporal Difference (TD) Loss




At $t = i - 1$,

1. Teacher-student model forecasts the next step, i.e. $\hat{X}^{(i-1)}{}_{i:i+H-1}$

At $t = i$,

1. Ground truth $\boldsymbol{x}_i$ arrives. The latest input sequence is $X_{i-L+1:i}$.
2. Freeze parameters of the teacher model $\boldsymbol{\theta}^{(T)}$
3. Teacher model predicts $\hat{X}^{(T,i)}{}_{i+1:i+H}$
4. Construct the pseudo label $\tilde{X}^{(i)}_{i:i+H-1} = [\boldsymbol{x}_i, \hat{X}^{(T,i)}_{i+1:i+H-1}]$
5. Compute the temporal difference loss
6. Update the parameters of student model $\boldsymbol{\theta}^{(S)}$

# Results

Integrate batch learning methods into the DSOF framework

Table 2: Comparison of MSE results between our framework and batch learning across various datasets and backbones. Our method integrates a MLP as the student model built on top of the teacher model, while batch learning relies exclusively on the teacher model. The better results from the two settings are highlighted in bold.

| | | DLinear | | FITS | | FSNet | | OneNet | | iTrans. | | PatchTST | | NSTrans. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H | Batch Learn. | DSOF | Batch Learn. | DSOF | Batch Learn. | DSOF | Batch Learn. | DSOF | Batch Learn. | DSOF | Batch Learn. | DSOF | Batch Learn. | DSOF |
| ECL | 1 | 2.842 | **2.065** | 2.870 | **2.235** | 3.8e+2 | **2.330** | 3.2e+1 | **4.733** | **1.976** | 2.430 | 4.770 | **2.244** | 3.9e+1 | **2.703** |
| | 24 | 1.5e+1 | **4.737** | **4.509** | 4.597 | 4.7e+2 | **5.475** | 8.5e+1 | **4.510** | **4.119** | 5.155 | 1.7e+1 | **5.169** | 4.0e+1 | **8.668** |
| | 48 | 2.7e+1 | **6.181** | **5.257** | 5.433 | 4.8e+2 | **7.000** | 1.5e+2 | **5.943** | **4.936** | 6.015 | 1.7e+1 | **6.665** | 3.8e+1 | **9.981** |
| ETTh2 | 1 | 0.470 | **0.365** | 0.522 | **0.375** | 1.1e+1 | **0.431** | 2.531 | **0.548** | 0.872 | **0.384** | 0.915 | **0.382** | 3.088 | **0.415** |
| | 24 | 2.269 | **1.701** | 2.189 | **1.757** | 1.9e+1 | **3.114** | 7.017 | **2.363** | 2.688 | **1.869** | 5.213 | **1.925** | 3.973 | **2.127** |
| | 48 | 3.389 | **3.082** | 3.275 | **2.988** | 2.3e+1 | **5.318** | 9.790 | **4.037** | 3.769 | **3.465** | 6.566 | **4.473** | 4.840 | **4.729** |
| ETTm1 | 1 | 0.112 | **0.105** | 0.123 | **0.111** | 0.190 | **0.121** | 0.156 | **0.096** | 0.179 | **0.152** | 0.155 | **0.108** | 1.101 | **0.141** |
| | 24 | 0.628 | **0.525** | 0.732 | **0.542** | 1.500 | **0.609** | 1.094 | **0.418** | 1.049 | **0.618** | 1.170 | **0.582** | 1.503 | **0.904** |
| | 48 | 0.818 | **0.695** | 0.900 | **0.716** | 2.279 | **0.843** | 1.588 | **0.554** | 1.320 | **0.856** | 2.103 | **0.805** | 1.547 | **1.119** |
| Ex. | 1 | **0.009** | **0.009** | 0.012 | **0.011** | 0.034 | **0.010** | 0.024 | **0.009** | 0.011 | **0.010** | 0.024 | **0.011** | 0.234 | **0.015** |
| | 24 | 0.098 | **0.095** | 0.095 | **0.093** | 0.754 | **0.120** | 0.487 | **0.152** | 0.117 | **0.110** | 0.443 | **0.103** | 0.335 | **0.127** |
| | 48 | 0.194 | **0.192** | 0.178 | **0.176** | 1.366 | **0.270** | 0.815 | **0.289** | **0.209** | 0.218 | 1.107 | **0.213** | 0.449 | **0.280** |
| Traffic | 1 | **0.302** | **0.302** | 0.342 | **0.313** | 0.599 | **0.228** | 0.265 | **0.264** | 0.243 | **0.242** | 0.280 | **0.232** | 0.762 | **0.257** |
| | 24 | 0.649 | **0.608** | 0.617 | **0.607** | 0.761 | **0.366** | 0.576 | **0.327** | 0.459 | **0.422** | 0.573 | **0.432** | 1.102 | **0.565** |
| | 48 | 0.769 | **0.681** | 0.707 | **0.680** | 0.824 | **0.403** | 0.683 | **0.369** | 0.517 | **0.457** | 0.621 | **0.454** | 1.001 | **0.604** |
| Weather | 1 | 0.357 | **0.337** | 0.359 | **0.341** | 0.728 | **0.388** | 0.478 | **0.296** | 0.474 | **0.336** | 0.423 | **0.361** | 2.489 | **0.336** |
| | 24 | 1.182 | **1.043** | 1.236 | **1.086** | 1.760 | **1.020** | 1.377 | **0.671** | 1.492 | **1.019** | 1.494 | **0.877** | 2.881 | **1.037** |
| | 48 | 1.702 | **1.447** | 1.634 | **1.434** | 2.620 | **1.415** | 2.699 | **0.909** | 1.740 | **1.435** | 2.033 | **1.308** | 3.039 | **1.419** |

# Results

DSOF, with or without the student model, ranks among the top two online learning methods.

Table 3: Comparison of the MSE results between DGrad, i.e. the modified online training frameworks from earlier studies (Pham et al., 2023; Zhang et al., 2023), as well as TFCL (Aljundi et al., 2019) and DER++ (Buzzega et al., 2020). All configurations have been modified to prevent information leakage.

| Teacher Model | Student Model | Framework | ECL 1 | ECL 24 | ECL 48 | ETTh2 1 | ETTh2 24 | ETTh2 48 | Traffic 1 | Traffic 24 | Traffic 48 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DLinear | ✗ | DGrad | 2.187 | 9.954 | 1e+01 | 0.386 | 3.038 | 6.251 | 1e+15 | 1.189 | 1.461 |
| | ✗ | TFCL | 3.033 | 6.897 | 9.746 | 0.718 | 2.240 | 3.457 | 0.324 | 0.697 | 0.745 |
| | ✗ | DER++ | 2.172 | 5.369 | 7.339 | 0.385 | 1.986 | 3.168 | 0.300 | 0.623 | 0.709 |
| | ✗ | DSOF (w/o $\boldsymbol{\theta}^{(S)}$) | 2.066 | 4.759 | 6.250 | **0.365** | **1.694** | **3.015** | **0.299** | **0.603** | 0.685 |
| | MLP | DSOF | **2.065** | **4.737** | **6.181** | **0.365** | 1.701 | 3.082 | 0.302 | 0.608 | **0.681** |
| FSNet | ✗ | DGrad | 3.197 | 5e+01 | 5e+01 | 0.470 | 2.956 | 5.359 | 1e+18 | 0.470 | 0.574 |
| | ✗ | TFCL | 2.968 | 7.482 | 8.512 | 0.726 | 3.469 | 6.256 | 0.315 | 0.418 | 0.467 |
| | ✗ | DER++ | 2.902 | 1e+01 | 1e+01 | 0.432 | **2.726** | 5.239 | 0.286 | 0.411 | 0.449 |
| | ✗ | DSOF (w/o $\boldsymbol{\theta}^{(S)}$) | 2.567 | 5.611 | 7.130 | **0.427** | 2.944 | **4.846** | 0.265 | **0.364** | **0.394** |
| | MLP | DSOF | **2.330** | **5.475** | **7.000** | 0.431 | 3.114 | 5.318 | **0.228** | 0.366 | 0.403 |
| PatchTST | ✗ | DGrad | 3.863 | 9.453 | 8.607 | 0.392 | 2.759 | 5.721 | 0.251 | 0.446 | 0.480 |
| | ✗ | TFCL | 3.805 | 7.653 | 1e+01 | 0.714 | 3.275 | 7.763 | 0.272 | 0.558 | 0.607 |
| | ✗ | DER++ | 3.403 | 6.846 | 8.025 | 0.386 | 2.163 | 5.849 | 0.242 | 0.478 | 0.502 |
| | ✗ | DSOF (w/o $\boldsymbol{\theta}^{(S)}$) | 3e+03 | 5.253 | **6.427** | **0.373** | **1.838** | **3.568** | **0.231** | 0.434 | 0.459 |
| | MLP | DSOF | **2.244** | **5.169** | 6.665 | 0.382 | 1.925 | 4.473 | 0.232 | **0.432** | **0.454** |

# Conclusion

- Problems in existing OTSF works
  - Information leakage
  - Only allow to forecast one future step
- Challenges of the redefined setting
  - Ground truth for the most recent prediction is not fully accessible at the subsequent time step
- Dual-Stream Online Forecasting (DSOF)
  - Slow stream (complete information): Updates the model with experience replay
  - Fast stream (immediate information): Updates the model with temporal difference loss