# Causal Generative Flows for Interventional and Counterfactual Time Series Prediction

Speaker: **Dongze Wu**

Georgia Tech

# Authors


Dongze Wu

Georgia Tech

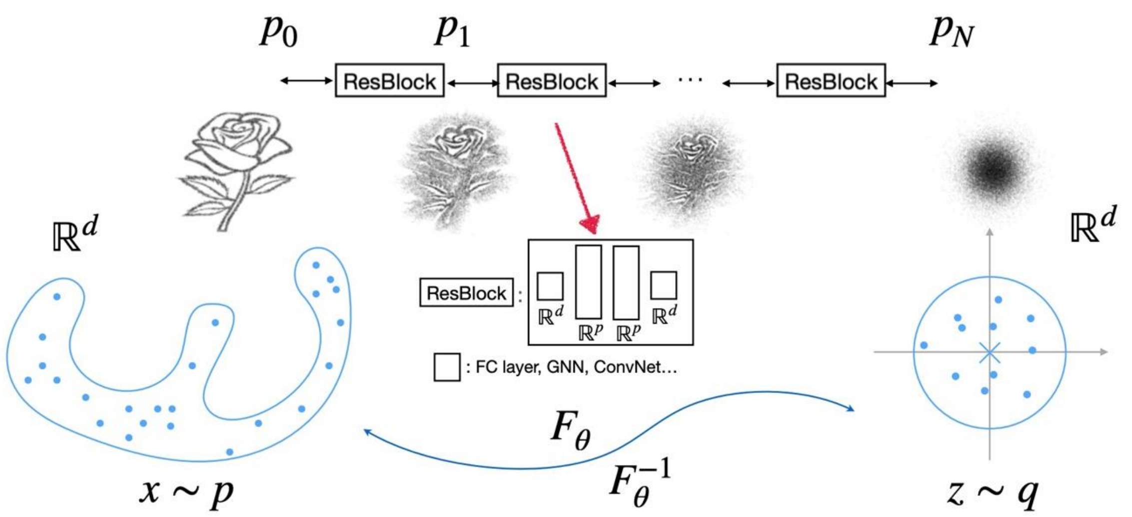
Feng Qiu

Argonne National Lab

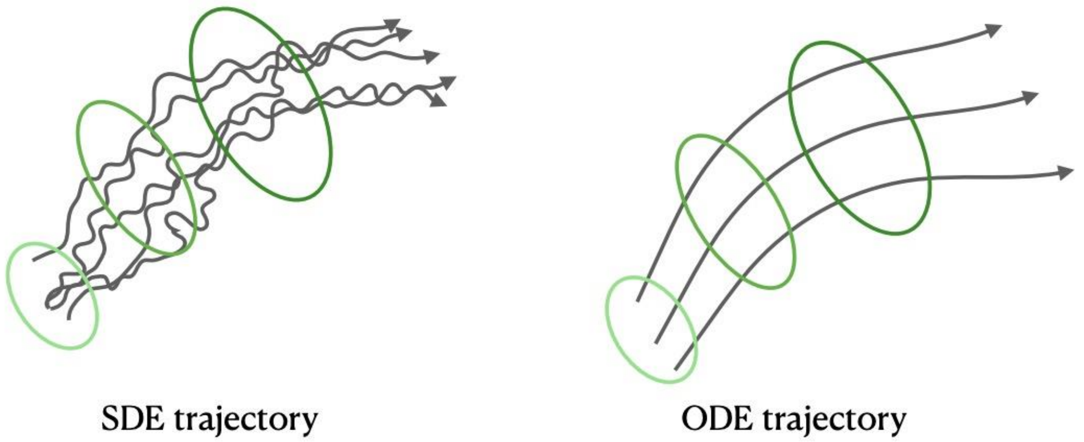
Dr. Yao Xie

Georgia Tech

# *Preliminary*: Flow Generative Models for Statistical Inferences




** Illustrative Figure on Traditional Flow Generative Model*




** Illustrative Figures comparing Diffusion v.s. Flow*

- Flow Model (ODE):

$$\frac{dx(t)}{dt} = v(x(t), t)$$

$$\partial_t \rho(x,t) + \nabla \cdot (\rho(x,t) v(x,t)) = 0$$

- Diffusion Model (SDE):

$$dx(t) = -\nabla v(x(t), t)dt + \sqrt{2} dW_t$$

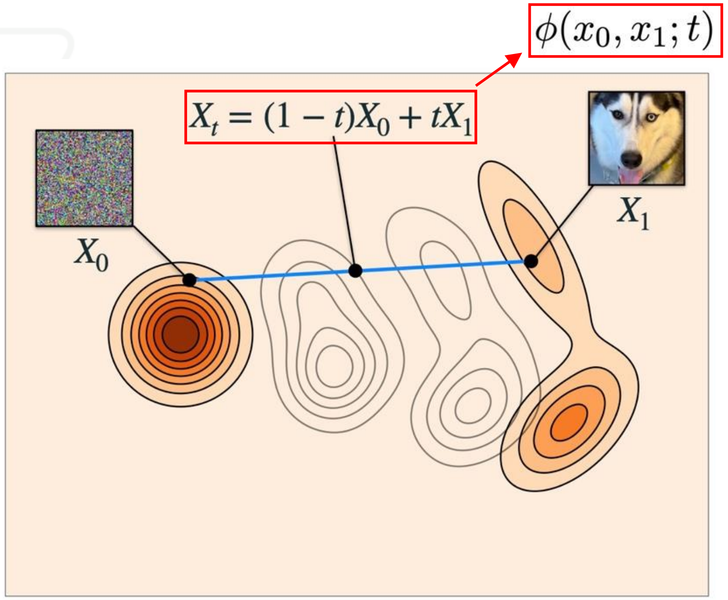
$$\partial_t \rho = \nabla \cdot (\rho_t \nabla v + \nabla \rho_t)$$

- Main Difference:

| Aspect | **Diffusion** | **Flow** |
|---|---|---|
| Generative Performance | Excellent | Fair |
| Statistical Inference | Poor | Excellent |
| Training Time | Poor | Good |

Georgia Tech

# Preliminary: Flow Matching for Generative Modeling (Lipman et al. (2023))




- Continuous Normalizing Flow:

$$\frac{dx(t)}{dt} = v(x(t), t)$$

$$\partial_t \rho(x,t) + \nabla \cdot (\rho(x,t) v(x,t)) = 0$$

Flow Matching:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t \sim \mathcal{U}[0,1],\, x \sim p(\cdot,t)} \left[ \| v(x(t),t) - u(x(t),t) \|^2 \right]$$

Conditional Flow Matching:

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{t \sim \mathcal{U}[0,1],\, x_0 \sim p(\cdot,0),\, x_1 \sim q(\cdot)} \left[ \left\| v(\phi,t) - \frac{d\phi}{dt} \right\|^2 \right]$$

# Existing Research on Causal Time Series

- Treatment Effects on Time Series

$$\tau_t = E[Y_t \mid A_{t-1} = j] - E[Y_t \mid A_{t-1} = k]$$

- Methods: conditional time-series forecasting (GPs, classical methods, transformer, etc.)

- Counterfactual Explainability

- Such works focus on *Interpretability*

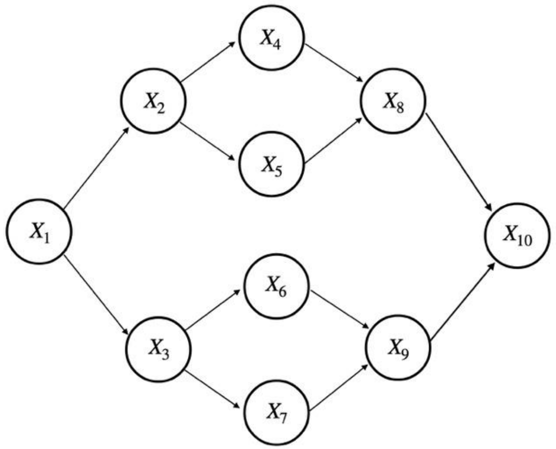
e.g., “What adjustments to a patient’s breathing signal would lead the model to forecast deeper sleep stages?”

- Methods: Optimization-based perturbation

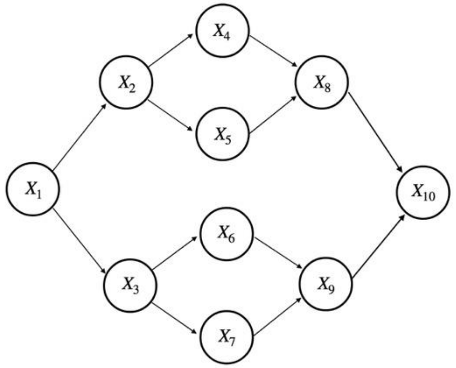
- Causal Discovery

- Inferring causal directed acyclic graph (DAG) from observed time series.

- Methods: Optimization over linear model



# Complimentary to Above Works




- Interventional and Counterfactual Forecasting
  - ➢ Assuming a known causal DAG
  - ➢ Enabling interventions on individual nodes at arbitrary times, and yielding coherent interventional and counterfactual forecasts of system-wide trajectories
  - ➢ Intervention:
    - o How an adjustment of turbine flow over a given time interval will influence the downstream time series signals over the causal DAG?

$$p(\mathbf{X}_{\tau+1:T} | \mathbf{x}_{1:\tau}, \mathrm{do}(X_{\mathcal{I}} := \gamma_{\mathcal{I}})) \qquad \mathcal{I} \subseteq [K] \times \{\tau+1, \ldots, T\}$$

  - ➢ Counterfactual:
    - o What would the future have looked like if we had set variable(s) $X_I$ to other values during the forecasting window?

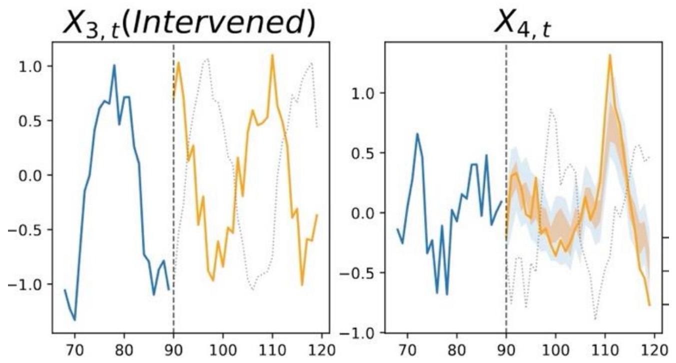
$$p(\mathbf{X}^{\mathrm{CF}}_{\tau+1:T} | \mathbf{x}_{1:\tau}, \mathbf{x}^{\mathrm{F}}_{\tau+1:T}, \mathrm{do}(X_{\mathcal{I}} := \gamma_{\mathcal{I}})) \quad \mathcal{I} \subseteq [K] \times \{\tau+1, \ldots, T\}$$

Georgia Tech

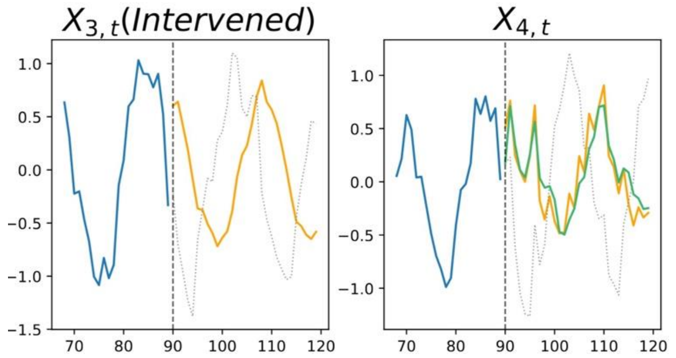
# Complimentary to Above Works

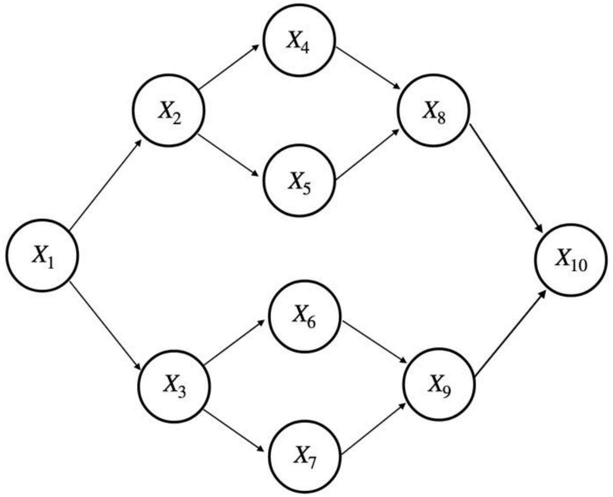
- Interventional and Counterfactual Prediction

Int.




CF.




### Intervention:

Conditional generation over causal DAG,

e.g., $\boldsymbol{p(\,X_8 \mid X_1 = \tau\,)}$

### Counterfactual:

Conditioned on the observed factual outcome, what would have occurred had we set the parent variables to different values?

e.g., $\boldsymbol{p\left(\,X_8^{CF} \mid X_8^{F}, X_1 = \tau\,\right)}$

### Most Common Counterfactual Inference (Static Data):

Assume a structural causal model (SCM): $X = f\left(X_{pa}, U\right)$

1. Abduction: Infer noise $U$ given factual data $X, X_{pa}$ and learned SCM $f^*$
2. Action: Set the intervened nodes to desired actions, i.e., $do\left(X_{\mathrm{pa}(i)} = \gamma\right)$
3. Prediction: Predict $X^{CF} = f^*(\gamma, U)$

Georgia Tech

# Settings and Goals

- A multivariate time series evolving over a causal DAG
- Nodes {1,…,K} in topologically sorted order
- $\mathbf{X_t} = \{X_{1,t}, \dots, X_{K,t}\}$ $\quad X_{pa(i),t} = \{X_{j,t} : j \in pa(i)\}$
- Context window: $\{\mathbf{X_1}, \dots, \mathbf{X_\tau}\}$;
- Forecasting window: $\{\mathbf{X}_{\tau+1}, \dots, \mathbf{X}_\mathrm{T}\}$

- Observational forecasting:

$$p(\mathbf{X}_{\tau+1:T} \mid \mathbf{x}_{1:\tau})$$

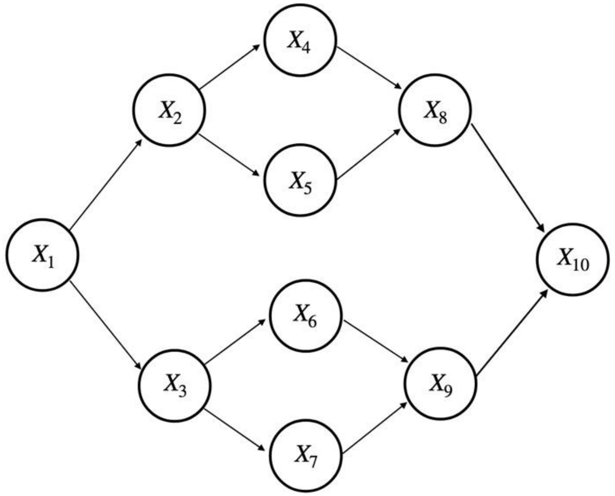
- Intervention Schedule: $\mathrm{I} \subseteq [K] \times \{\tau + 1, \dots, T\}$

- Interventional Forecasting:

$$p(\mathbf{X}_{\tau+1:T} | \mathbf{x}_{1:\tau}, \mathrm{do}(X_{\mathcal{I}} := \gamma_{\mathcal{I}})) \quad \mathcal{I} \subseteq [K] \times \{\tau+1, \dots, T\}$$

- Counterfactual Forecasting:

$$p(\mathbf{X}^{\mathrm{CF}}_{\tau+1:T} | \mathbf{x}_{1:\tau}, \mathbf{x}^{\mathrm{F}}_{\tau+1:T}, \mathrm{do}(X_{\mathcal{I}} := \gamma_{\mathcal{I}})) \quad \mathcal{I} \subseteq [K] \times \{\tau+1, \dots, T\}$$




Georgia Tech

# Time-Conditioned Continuous Normalizing Flow

- Hidden State Conditioning:

$$h_{i,t} = \mathrm{RNN}(\mathrm{concat}\{x_{i,t}, c_{i,t}\}, h_{i,t-1})$$

$$H_{i,t-1} := (h_{i,t-1}, h_{\mathrm{pa}(i),t-1})$$

- Neural ODE of the Time-Conditioned CNF:

$$\frac{dx_t(s)}{ds} = v(x_t(s), s; H_{t-1}), \quad s \in [0,1], \quad t \in \{\tau+1, \tau+2, \dots, T\}$$

- Training Loss (Flow Matching):

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\mathbf{x}_{1:T} \sim p_{\mathcal{X}}} \left[ \frac{1}{K(T-\tau)} \sum_{i=1}^{K} \sum_{t=\tau+1}^{T} \mathbb{E}_{s \sim \mathcal{U}[0,1],\, z \sim \mathcal{N}(0,I)} \left\| v\big(\phi(x_{i,t}, z; s),\ s;\ H_{i,t-1}\big) - \partial_s \phi(x_{i,t}, z; s) \right\|_2^2 \right]$$

# Time-Conditioned Continuous Normalizing Flow

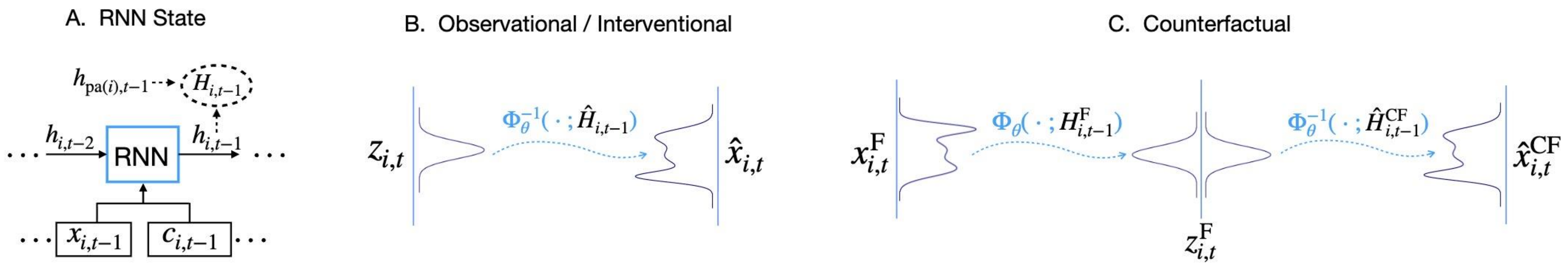



Figure 1: **(A)** RNN State Update. **(B)** Observational/Interventional Forecasting. Forecasts are generated by decoding from latent $z_{i,t} \sim N(0,1)$, conditioned on $\hat{H}_{i,t-1}$ updated with the last predicted $\hat{x}_{i,t-1}$. **(C)** A factual observation $x^{\mathrm{F}}_{i,t}$ is encoded with its factual state $H^{\mathrm{F}}_{i,t}$ into $z^{\mathrm{F}}_{i,t}$, then decoded under the counterfactual state $\hat{H}^{\mathrm{CF}}_{i,t-1}$ to yield $\hat{x}^{\mathrm{CF}}_{i,t}$. Factual states $H^{\mathrm{F}}_{i,t-1}$ are updated from observed $x^{\mathrm{F}}_{i,t-1}$, while counterfactual states $\hat{H}^{\mathrm{CF}}_{i,t-1}$ are updated from the previously generated $\hat{x}^{\mathrm{CF}}_{i,t-1}$.

# Observational/Interventional Forecasting

**Algorithm 3:** Time Series Observational/Interventional Forecasting

**Input:** Context window $\{x_{i,1:\tau}\}_{i=1}^{K}$; intervention schedule $\mathcal{I}$ with values $\{\gamma_{i,t}\}$
Initialize hidden states $\hat{H}_{i,\tau} = H_{i,\tau}$ with $x_{i,1:\tau}$ for all $i = 1, \ldots, K$
**for** $t = \tau + 1$ **to** $T$ **do**
  **for** $i = 1, \ldots, K$ **do** {topological order}
    **if** $(i, t) \in \mathcal{I}$ **then**
      $\hat{x}_{i,t} \leftarrow \gamma_{i,t}$
    **else**
      Sample $z_{i,t} \sim \mathcal{N}(0, 1)$
      $\hat{x}_{i,t} \leftarrow \textsc{Dec}(z_{i,t}, \hat{H}_{i,t-1})$ {Alg. 1}
    **end if**
    $h_{i,t}, h_{\mathrm{pa}(i),t} \xleftarrow{\text{update}} (\hat{x}_{i,t},\ \hat{x}_{\mathrm{pa}(i),t})$
    $\hat{H}_{i,t} \leftarrow (h_{i,t}, h_{\mathrm{pa}(i),t})$
  **end for**
**end for**
**Output:** $\{\hat{x}_{i,t}\}_{i=1..K,\ t=\tau+1,..,T}$

**Algorithm 1:** $\textsc{Dec}(\cdot, \cdot)$

**Input:** Latent input $z_{i,t}$; conditioning hidden state $H_{i,t-1}$
Integrate the ODE backward from $s = 1$ to $s = 0$ with $x(1) = z_{i,t}$:

$$x(0) \leftarrow \Phi_\theta^{-1}(z_{i,t};\ H_{i,t-1}) = z_{i,t} - \int_0^1 v\big(x(s),\ s;\ H_{i,t-1}\big)\, ds$$

$\hat{x}_{i,t} \leftarrow x(0)$
**Return:** $\hat{x}_{i,t}$

† Empirically, we use Runge–Kutta numerical integration.

# Counterfactual Forecasting

**Algorithm 4:** Counterfactual Time Series Generation

```
Input: Context window $\{x_{i,1:\tau}\}_{i=1}^{K}$; factual sample $\{x^{\mathrm{F}}_{i,\tau+1:T}\}_{i=1}^{K}$; intervention schedule $\mathcal{I}$ with values $\{\gamma_{i,t}\}$
Obtain factual hidden states $\{H^{\mathrm{F}}_{i,t}\}_{t=\tau}^{T-1}$ from the context $\{x_{i,1:\tau}\}$ and observed factual $\{x^{\mathrm{F}}_{i,\tau+1:T}\}$
Initialize counterfactual hidden states $\hat{H}^{\mathrm{CF}}_{i,\tau} = H_{i,\tau}$ with context $\{x_{i,1:\tau}\}$ for all $i = 1, \dots, K$
for $t = \tau + 1$ to $T$ do
    for $i = 1, \dots, K$ do {nodes in topological order}
        if $(i, t) \in \mathcal{I}$ then
            $\hat{x}^{\mathrm{CF}}_{i,t} \leftarrow \gamma_{i,t}$
        else
            $z^{\mathrm{F}}_{i,t} \leftarrow \mathrm{ENC}(x^{\mathrm{F}}_{i,t}, H^{\mathrm{F}}_{i,t-1})$        {Algorithm 2: Abduction}
            $\hat{x}^{\mathrm{CF}}_{i,t} \leftarrow \mathrm{DEC}(z^{\mathrm{F}}_{i,t}, \hat{H}^{\mathrm{CF}}_{i,t-1})$     {Algorithm 1: Action-Prediction}
        end if
        $h_{i,t}, h_{\mathrm{pa}(i),t} \xleftarrow{\text{update}} (\hat{x}^{\mathrm{CF}}_{i,t},\ \hat{x}^{\mathrm{CF}}_{\mathrm{pa}(i),t})$
        $\hat{H}^{\mathrm{CF}}_{i,t} \leftarrow (h_{i,t}, h_{\mathrm{pa}(i),t})$
    end for
end for
Output: $\{\hat{x}^{\mathrm{CF}}_{i,t}\}_{i=1..K,\ t=\tau+1,..T}$
```

**Algorithm 2:** $\mathrm{ENC}(\cdot,\cdot)$

1. **Input:** Observed factual value $x^{\mathrm{F}}_{i,t}$; conditioning factual state $H^{\mathrm{F}}_{i,t-1}$
2. Integrate the ODE forward from $s = 0$ to $s = 1$ with $x(0) = x^{\mathrm{F}}_{i,t}$:

$$x(1) \leftarrow \Phi_\theta(x^{\mathrm{F}}_{i,t};\ H^{\mathrm{F}}_{i,t-1}) = x^{\mathrm{F}}_{i,t} + \int_0^1 v\big(x(s),\ s;\ H^{\mathrm{F}}_{i,t-1}\big)\,ds$$

3. $z^{\mathrm{F}}_{i,t} \leftarrow x(1)$
4. **Return:** $z^{\mathrm{F}}_{i,t}$

† Empirically, we use Runge–Kutta numerical integration.

**Algorithm 1:** $\mathrm{DEC}(\cdot,\cdot)$

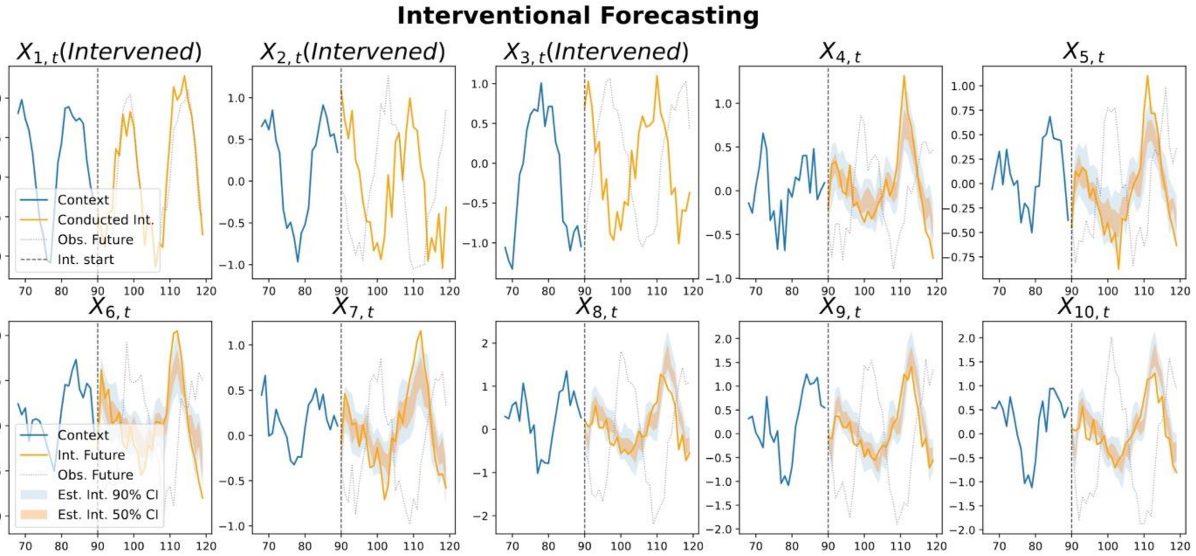
1. **Input:** Latent input $z_{i,t}$; conditioning hidden state $H_{i,t-1}$
2. Integrate the ODE backward from $s = 1$ to $s = 0$ with $x(1) = z_{i,t}$:

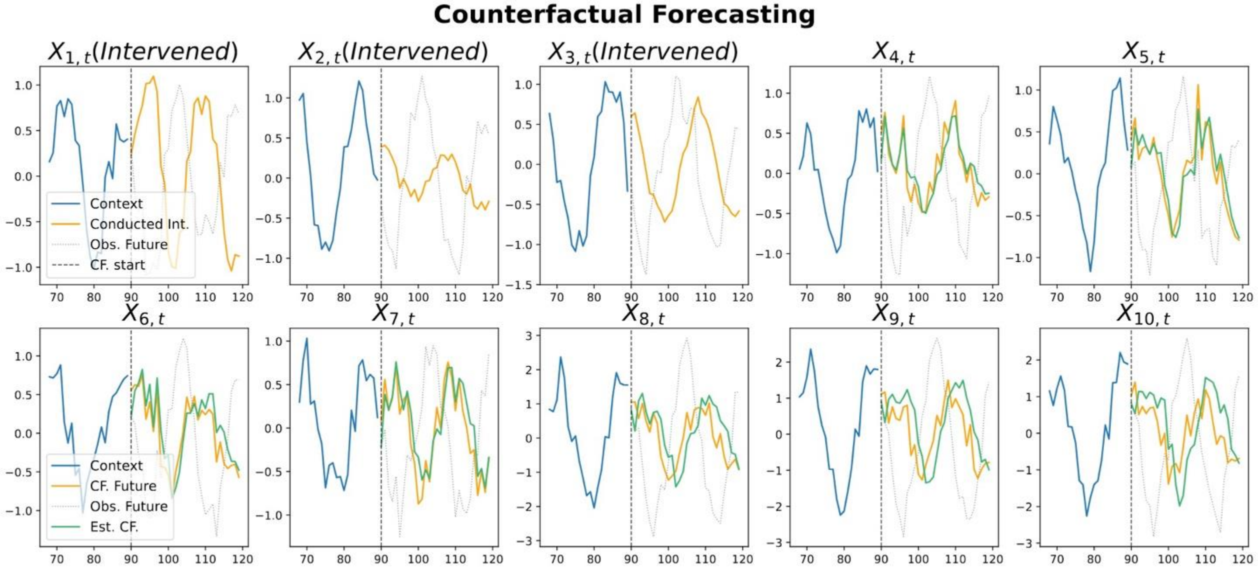
$$x(0) \leftarrow \Phi_\theta^{-1}(z_{i,t};\ H_{i,t-1}) = z_{i,t} - \int_0^1 v\big(x(s),\ s;\ H_{i,t-1}\big)\,ds$$

3. $\hat{x}_{i,t} \leftarrow x(0)$
4. **Return:** $\hat{x}_{i,t}$

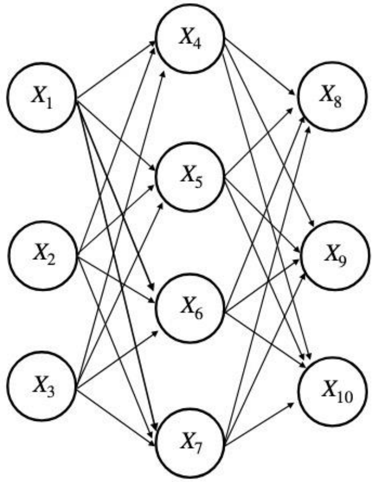
† Empirically, we use Runge–Kutta numerical integration.

# Interventional & Counterfactual Illustrations




$$p(\mathbf{X}_{\tau+1:T} | \mathbf{x}_{1:\tau}, \mathrm{do}(X_{\mathcal{I}} := \gamma_{\mathcal{I}}))$$




$$p(\mathbf{X}^{\mathrm{CF}}_{\tau+1:T} | \mathbf{x}_{1:\tau}, \mathbf{x}^{\mathrm{F}}_{\tau+1:T}, \mathrm{do}(X_{\mathcal{I}} := \gamma_{\mathcal{I}}))$$

Georgia Tech

# Counterfactual Recovery Properties

We assume that the structural causal models (SCM) is given by:

$$X_t \coloneqq f\left(X_{<t}, X_{pa,<t}, U_t\right)$$

**Assumption 5.1.**
(A1) $U_t \perp\!\!\!\perp (X_{<t}, X_{\mathrm{pa},<t})$.
(A2) The structural causal equation $f(\cdot, U_t)$ is monotone in $U_t$.
(A3) For the encoded latent variable $Z_t = \Phi_\theta(X_t; H_{t-1})$, the conditional distribution satisfies $p_\theta(Z_t \mid H_{t-1}) = q(Z_t) = N(Z_t; 0, 1)$.

**Proposition 5.3** (Encoded as a function of the exogenous noise $U_t$). *Let Assumption 5.1 hold. Without loss of generality, suppose the exogenous noise $U_t \sim \mathrm{Unif}[0,1]$. At each time t, the observed variable is generated by the structural causal model $X_t = f(X_{<t}, X_{\mathrm{pa},<t}, U_t)$, and that the flow encoder produces $Z_t = \Phi_\theta(X_t; H_{t-1})$. Then there exists a continuously differentiable bijection $g : \mathcal{U} \to \mathcal{Z}$, functionally invariant to $H_{t-1}$, such that,*

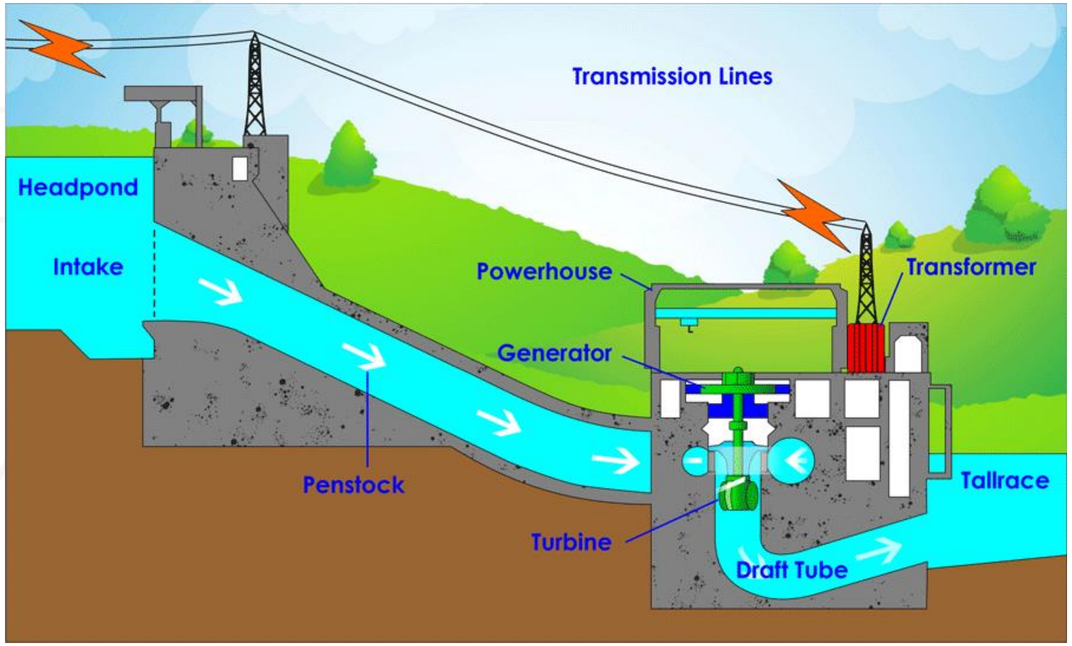
$$Z_t = \Phi_\theta\big(X_t; H_{t-1}\big) = \Phi_\theta\big(f(X_{<t}, X_{\mathrm{pa},<t}, U_t); H_{t-1}\big) = g\big(U_t\big) \quad a.s. \tag{16}$$

**Corollary 5.5** (Counterfactual recovery). *Let Assumption 5.1 hold. Consider a factual sample generated by the structural causal model $X_t^{\mathrm{F}} = f(X_{<t}, X_{\mathrm{pa},<t}, U_t)$, and let its encoded latent be $Z_t^{\mathrm{F}} \coloneqq \Phi_\theta\big(X_t^{\mathrm{F}}; H_{t-1}^{\mathrm{F}}\big)$. At time step t, we apply the intervention $\mathrm{do}\big(X_{<t} = \hat{X}_{<t}^{\mathrm{CF}}, X_{\mathrm{pa},<t} = \hat{X}_{\mathrm{pa},<t}^{\mathrm{CF}}\big)$, yielding the counterfactual hidden state $\hat{H}_{t-1}^{\mathrm{CF}}$. Then the decoder recovers the true counterfactual at time step t almost surely:*

$$\hat{X}_t^{\mathrm{CF}} \coloneqq \Phi_\theta^{-1}\big(Z_t^{\mathrm{F}}; \hat{H}_{t-1}^{\mathrm{CF}}\big) = X_t^{\mathrm{CF}}.$$

Georgia Tech

# Argonne Hydropower System



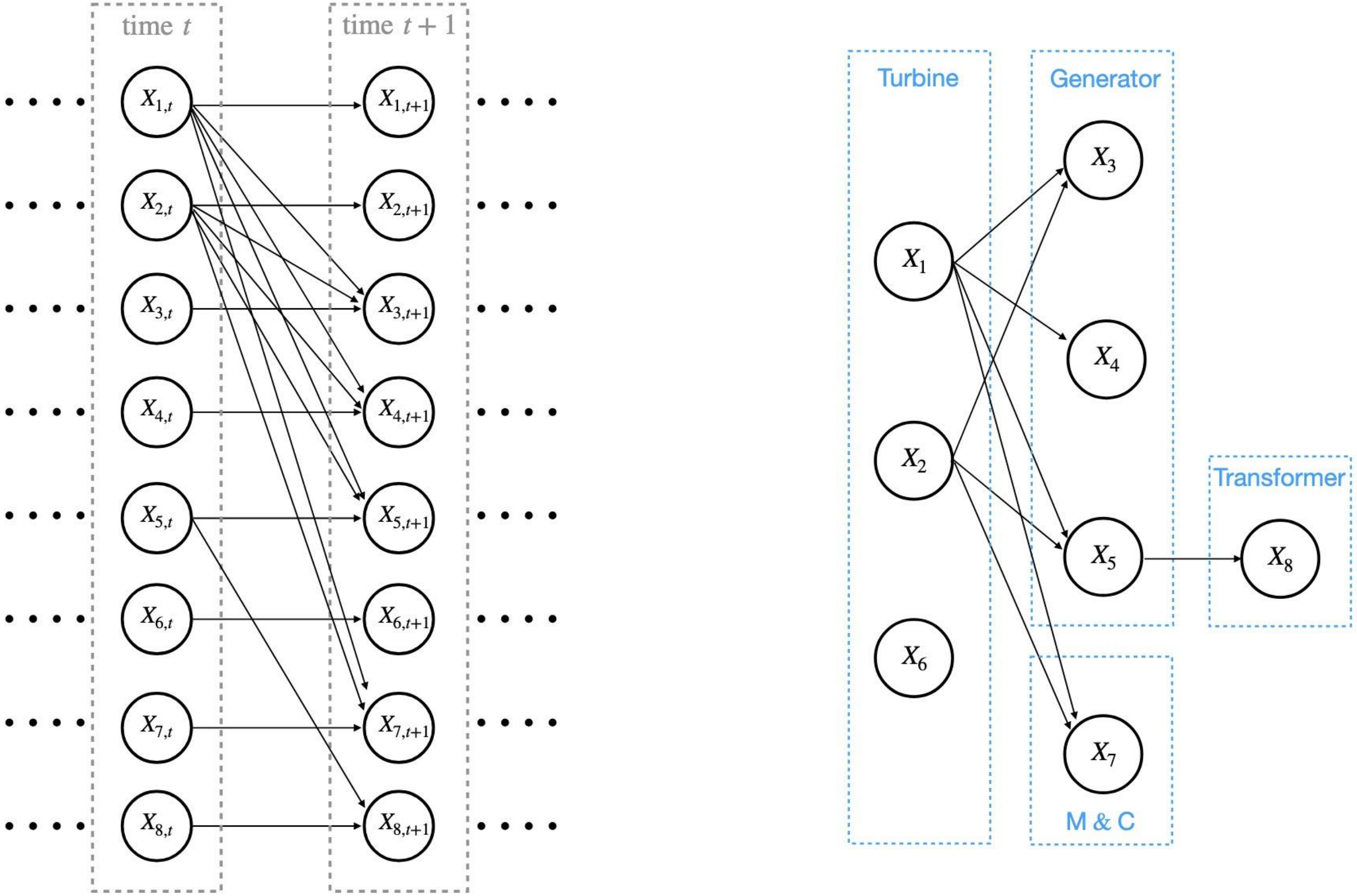
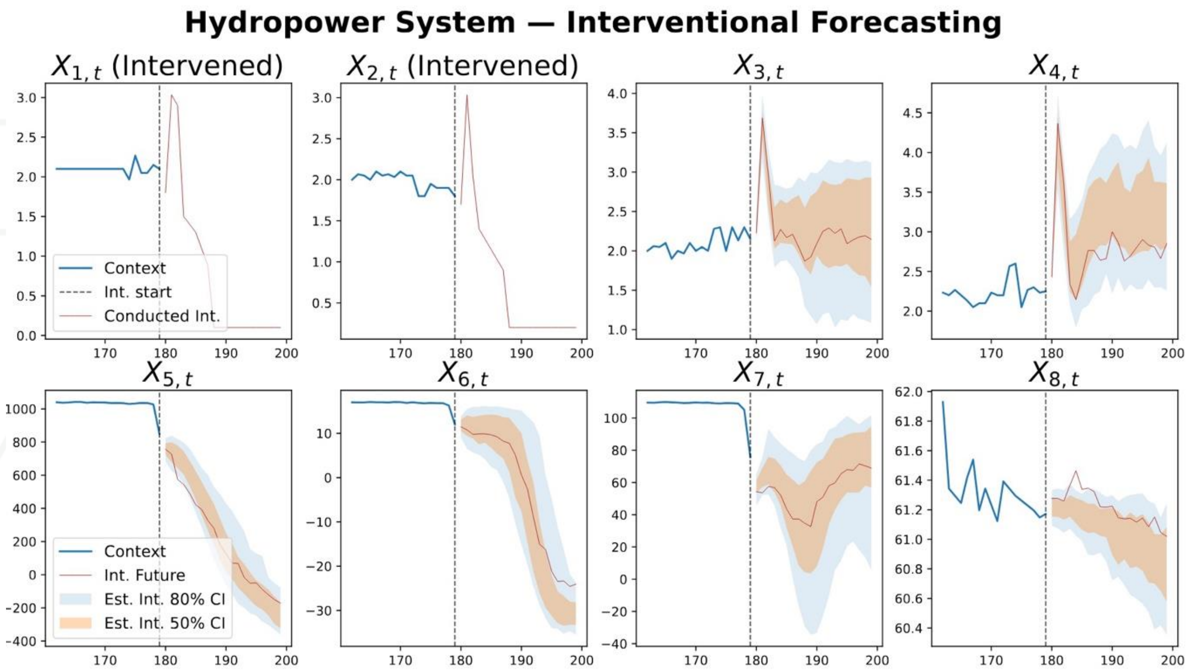




Figure 11: **Hydropower** system graph over 8 nodes. Exogenous variables $U_{i,t}$ are omitted for clarity but exist for every node at each time $t$. **Left:** Full node-level causal structure between consecutive time, with all variables $\{X_{1,t}, \ldots, X_{8,t}\}$ present at each step. **Right:** Rolled-up (time-suppressed) view over different nodes $\{X_1, \ldots, X_8\}$. Each arrow $X_i \to X_j$ (with $i \neq j$) denotes a lag-1 temporal dependency $X_{i,t-1} \to X_{j,t}$ that holds for all $t$. Both panels depict the same underlying structure.

Georgia Tech

# Hydropower – Interventional Forecasting




| | Hydropower System | |
|---|---|---|
| | Obs. | Int. |
| **DoFlow** | $\mathbf{1.13}_{\pm .18}$ | $\mathbf{1.21}_{\pm .19}$ |
| GRU | $2.05_{\pm .32}$ | $2.45_{\pm .35}$ |
| TFT | $1.82_{\pm .25}$ | $2.16_{\pm .41}$ |
| TiDE | $1.49_{\pm .24}$ | $2.08_{\pm .40}$ |
| TSMixer | $1.51_{\pm .25}$ | $2.11_{\pm .32}$ |
| DeepVAR | $1.78_{\pm .26}$ | $2.39_{\pm .28}$ |
| MQF2 | $1.97_{\pm .24}$ | $2.62_{\pm .34}$ |

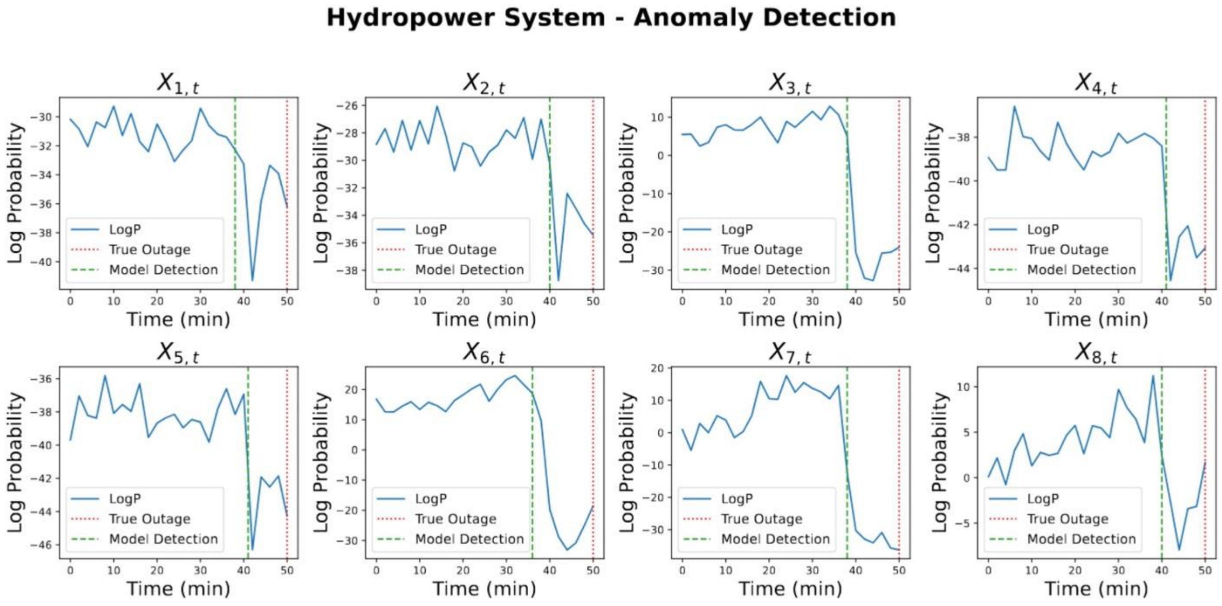
Table 7: RMSE for observational and interventional time-series forecasting in the hydropower system.

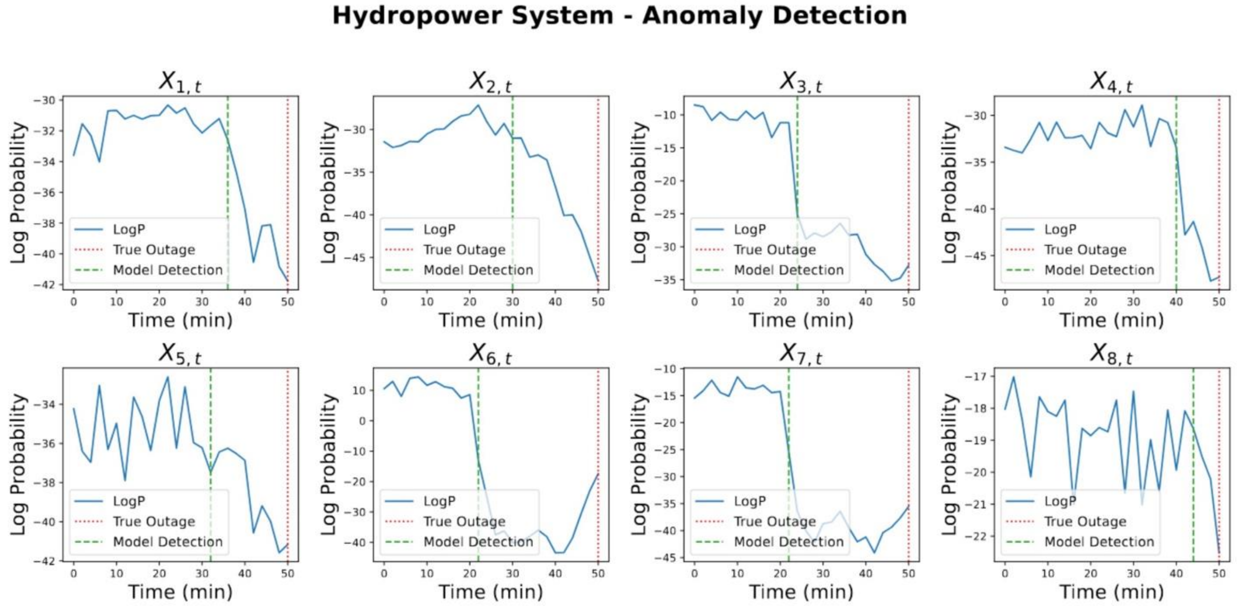
Georgia Tech

# Hydropower – Anomaly Detection

**Proposition 4.1** *Given base samples $z_{\tau+1:T} \sim q(\cdot)$, the log-density of the generated time series obtained via the continuous normalizing flow is:*

$$\log p_{\theta, X_{\tau+1:T}}\left(\hat{x}_{\tau+1:T} \mid \hat{H}_{\tau}, z_{\tau+1:T}\right) = \sum_{t=\tau+1}^{T} \left[\log q(z_t) + \int_0^1 \nabla \cdot\, v_\theta\left(x_t(s), s;\, \hat{H}_{t-1}\right) ds\right].$$



# Thank you!

Georgia Tech