# Evaluation of Information-Theoretic Measures in Echo State Networks on the Edge of Stability

Miloslav Torda

Supervisor: prof. Ing. Igor Farkaš, Dr.

# Previous results

**Artificial neural networks:**

- Task performance is maximized when the network is operating in a state near the edge of chaos, or stability (Bertschinger and Natschläger, 2004)
- Information measures are maximized near the edge of stability (Boedecker et. al., 2012)

**Biological neural networks:**

- Cortical circuits may be tuned to criticality for optimized behavior (Beggs, 2008)

# Echo state network:
## Model

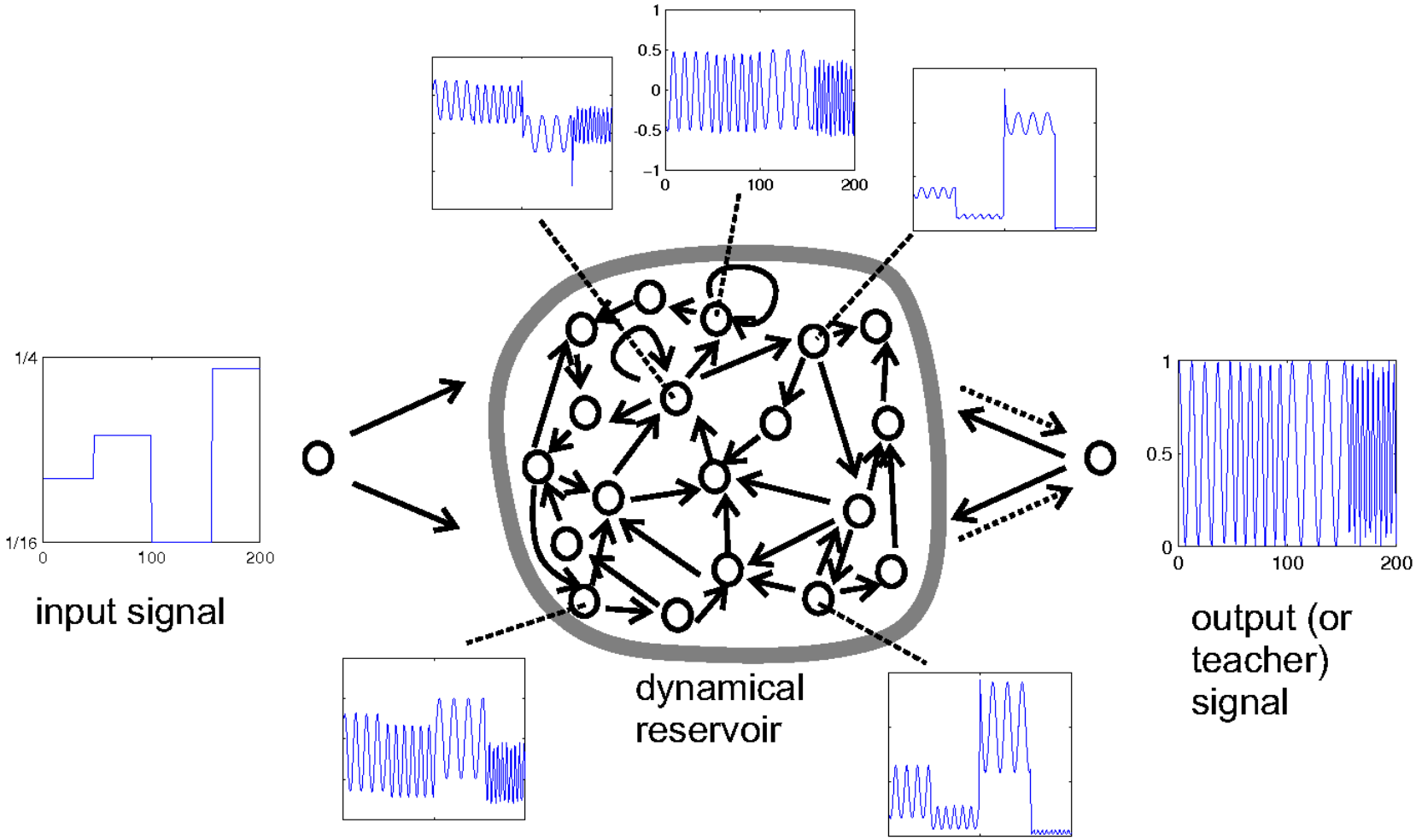



Figure: Herbert Jaeger (2007) Echo state network. Scholarpedia, 2(9):2330.

# Echo state network
## General model

**Update equations:**

$$\mathbf{x}(t) = \tanh(\mathbf{w}^{\text{in}} u(t) + \mathbf{W}\mathbf{x}(t-1))$$

$$\mathbf{x}(t) \in \mathbb{R}^{N_x}, \mathbf{w}^{\text{in}} \in \mathbb{R}^{N_x}, \mathbf{W} \in \mathbb{R}^{N_x \times N_x}$$

**Output layer:**

$$\mathbf{y}(t) = \mathbf{W}^{\text{out}} \mathbf{x}(t)$$

$$\mathbf{y}(t) \in \mathbb{R}^{N_y}, \mathbf{W}^{\text{out}} \in \mathbb{R}^{\mathbf{N_y} \times \mathbf{N_x}}$$

**Learning:**

$$\mathbf{W}^{\text{out}} = \mathbf{U}\mathbf{X}^{+}$$

$$\mathbf{X}^{+} = \mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}$$

# Information measures

**Transfer entropy:**

$$TE^{(k,l)}_{Y\to X} = I(X_t : \mathbf{Y}^{(l)}_{t-u} \mid \mathbf{X}^{(k)}_{t-1})$$

$$= \sum_{\mathbf{x}^{(k)}_{t-1},\mathbf{y}^{(l)}_{t-u}} p(\mathbf{x}^{(k)}_{t-1}, \mathbf{y}^{(l)}_{t-u}) \sum_{x_t} p(x_t \mid \mathbf{x}^{(k)}_{t-1}, \mathbf{y}^{(l)}_{t-u}) log_2 \frac{p(x_t \mid \mathbf{x}^{(k)}_{t-1}, \mathbf{y}^{(l)}_{t-u})}{p(x_t \mid \mathbf{x}^{(k)}_{t-1})}$$

**Active information storage:**

$$AIS^{(k)}_X = I(X_t : \mathbf{X}^{(k)}_{t-1})$$

$$= \sum_{x_t,\mathbf{x}^{(k)}_{t-1}} p(x_t, \mathbf{x}^{(k)}_{t-1}) log_2 \frac{p(x_t, \mathbf{x}^{(k)}_{t-1})}{p(x_t)(\mathbf{x}^{(k)}_{t-1})}$$

where

$$\mathbf{X}^{(k)}_{t-1} = (X_{t-1}, X_{t-1-\tau_k}, X_{t-1-2\tau_k}, \ldots, , X_{t-1-(k-1)\tau_k})$$

$$\mathbf{Y}^{(l)}_{t-u} = (Y_{t-u}, Y_{t-u-\tau_l}, Y_{t-u-2\tau_l}, \ldots, , Y_{t-u-(l-1)\tau_l})$$

**Transfer entropy:**

$$\widehat{\mathrm{TE}}_{Y\to X}^{(k,l)} = \Psi(K) + \left\langle \Psi(n_{\mathbf{x}_{t-1}^{(k)}} + 1) - \Psi(n_{x_t,\mathbf{x}_{t-1}^{(k)}} + 1) - \Psi(n_{\mathbf{x}_{t-1}^{(k)},\mathbf{y}_{t-1}^{(l)}} + 1) \right\rangle_t$$

where $\Psi$ denotes the digamma function, $n_{(\cdot)}$ denotes the number of nearest neighbors in $\epsilon$-hypercubes centered at $(\cdot)$ in the marginal spaces where $\epsilon$ is given by the Chebyshev distance of the realization $(x_t, \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-1}^{(l)})$ at time step $t$ to its $K$th nearest neighbor in the joint space, and $\langle\cdot\rangle_t$ denotes the time-average.

**Active information storage:**

$$\widehat{\mathrm{AIS}}_X^{(k)} = \Psi(K) + \Psi(N) - \left\langle \Psi(n_{\mathbf{x}_{t-1}^{(k)}} + 1) + \Psi(n_{x_t} + 1) \right\rangle_t$$

# Estimating criticality

## Maximum Lyapunov characteristic exponent

Let $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$ be a discrete time dynamical system and $\mathbf{x}'(0) = \mathbf{x}(0) + \delta\mathbf{x}(0)$ be the initial conditions a trajectory $\mathbf{x}'(t)$ obtained by an infinitesimal displacement from $\mathbf{x}(0)$ such that $\gamma_0 = \|\delta\mathbf{x}(0)\| \ll 1$. Then the maximum Lyapunov characteristic exponent is

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \ln\left(\frac{\gamma_t}{\gamma_0}\right) < \infty$$

where $\gamma_t = \|\mathbf{x}'(t) - \mathbf{x}(t)\|$.

$$\gamma_t \sim \gamma_0 e^{\lambda t} \Rightarrow \begin{cases} \lambda > 0 & \text{sensitive to initial conditions} \\ \lambda \approx 0 & \text{edge of criticality} \\ \lambda < 0 & \text{sub-critical} \end{cases}$$

# Experimental setup
## ESN model

- N=100 reservoir units
- single input unit
- output units
  - 120 for the MC task
  - single unit for the prediction task
- $w_j^{\text{in}} \in \mathcal{U}(-0.1; 0.1), \forall j = 1, \ldots, N$
- $w_{ij}^{\text{res}} \in \mathcal{N}(0; 0.5), \forall i, j = 1, \ldots, N$
- 100 samples transients
- 1000 samples training
- 2000 samples testing

**Prediction tasks:**

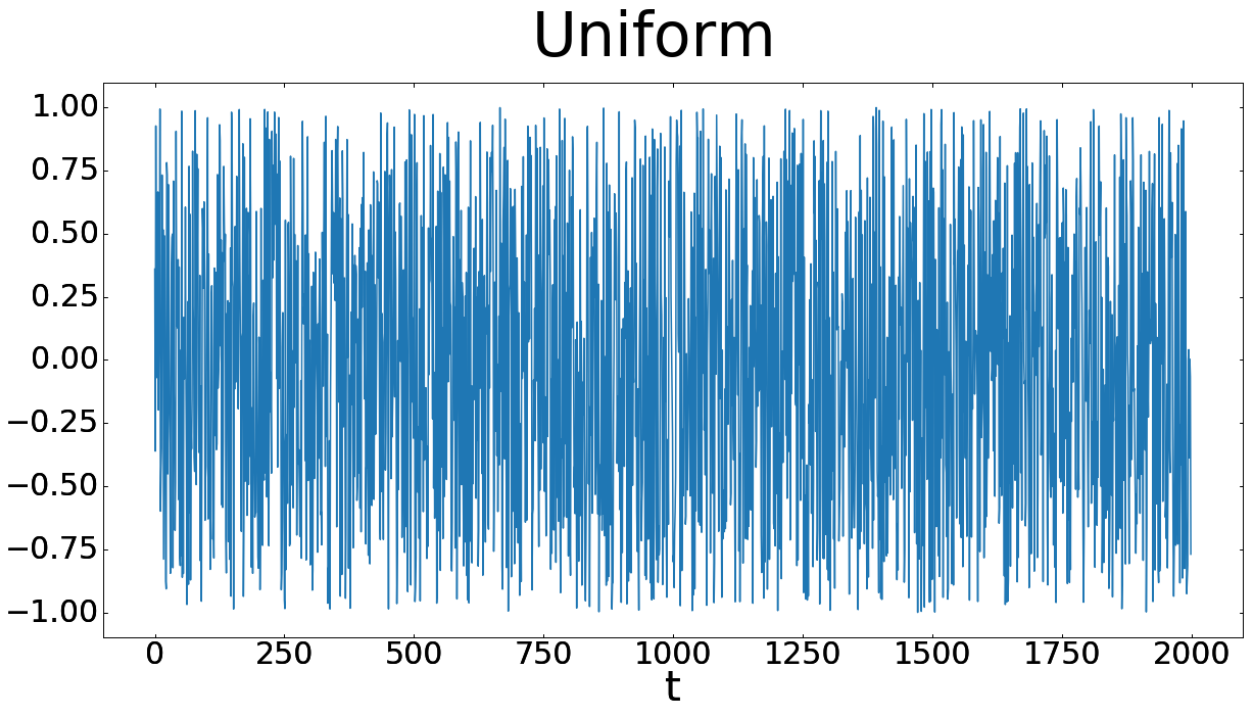
$$\mathrm{NRMSE} = \sqrt{\frac{\langle (\hat{y}(t) - y(t))^2 \rangle_t}{\langle (y(t) - \langle y(t) \rangle_t)^2 \rangle_t}}$$

**MC task:**

$$\mathrm{MC} = \sum_{q=1}^{q_{\max}} \mathrm{MC}_q = \sum_{q=1}^{q_{\max}} \frac{\mathrm{cov}^2(u(t-q), y_q(t))}{\mathrm{var}(u(t)) \cdot \mathrm{var}(y_q(t))}$$

# Benchmark tasks
## MC

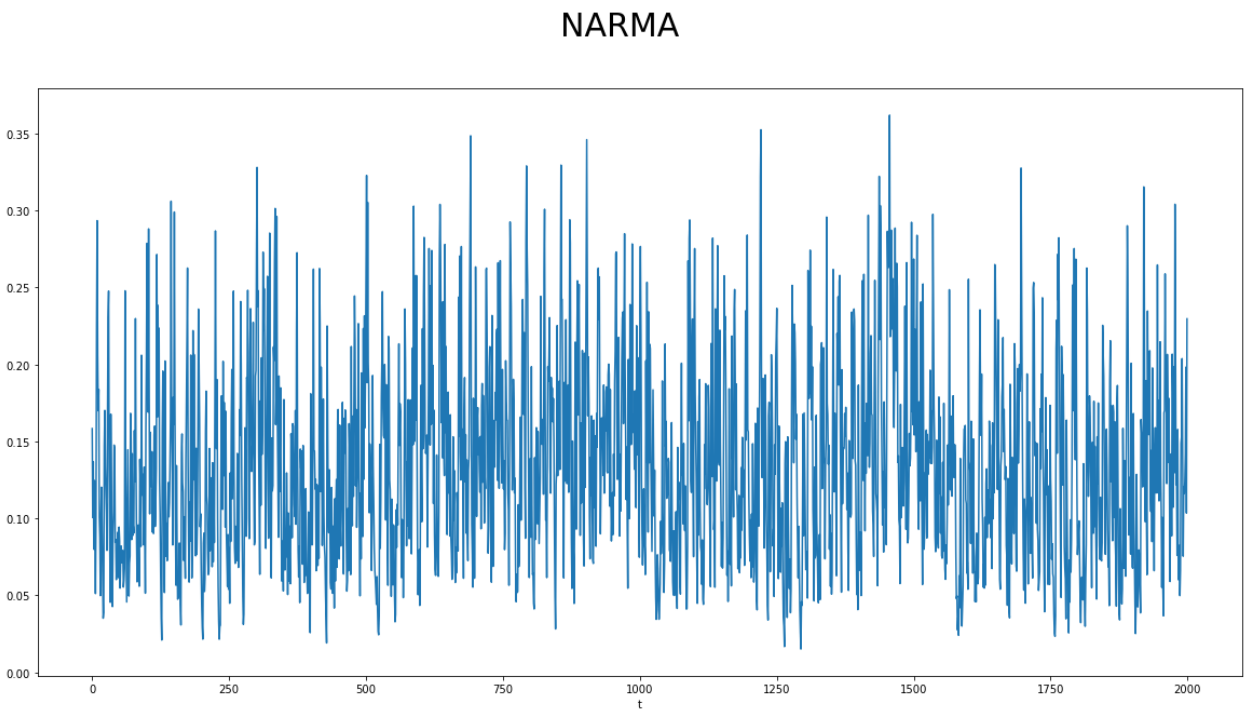
$$\forall\, t : u(t) \sim \mathcal{U}(-1; 1)$$



# Benchmark tasks
## NARMA

$$u(t+1) = 0.2\,u(t) + 0.004\,u(t)\sum_{i=0}^{29} u(t-i) + 1.5\,q(t-29)q(t) + 0.001$$

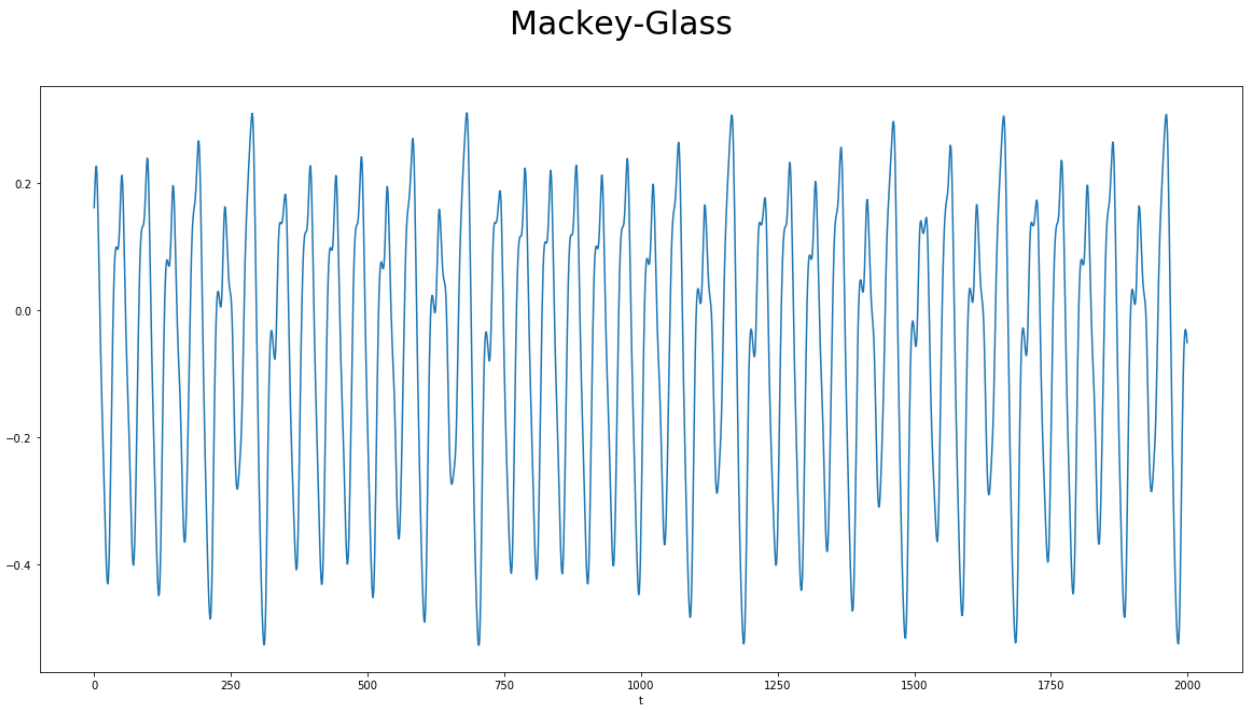
where $\forall\, t : q(t) \sim \mathcal{U}(0; 0.5)$



# Benchmark tasks
## Mackey-Glass system

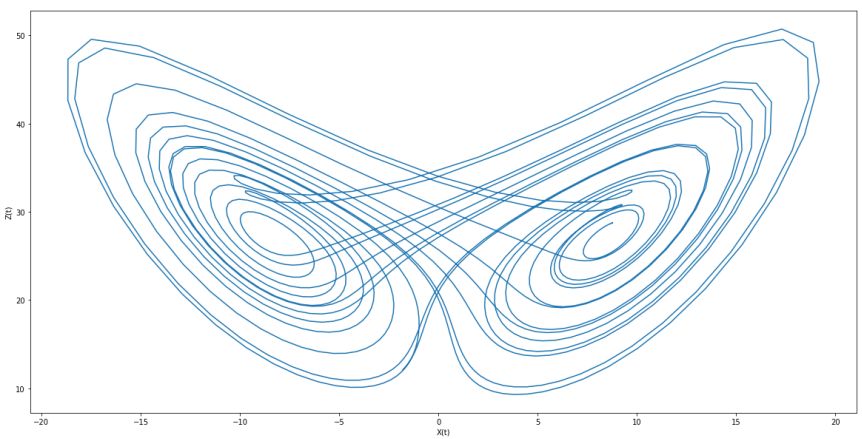
$$\frac{dy}{dt} = 0.2\frac{y_{17}}{1 + y_{17}^{10}} - 0.1y$$

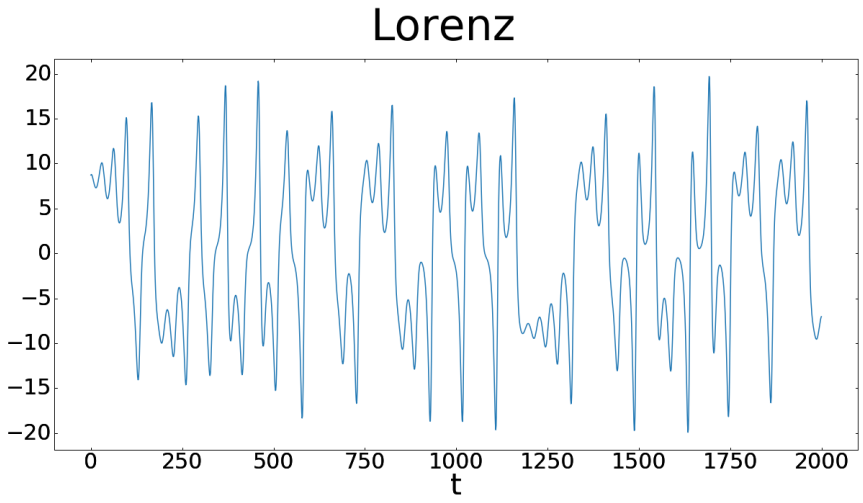
where $u_{17}$ is the value of $u$ at time $t - 17$.



# Benchmark tasks
## Lorenz system

$$\frac{dx}{dt} = 10(y - x)$$
$$\frac{dy}{dt} = x(28 - z) - y$$
$$\frac{dz}{dt} = xy - \frac{8}{3}z$$



**ESN:**

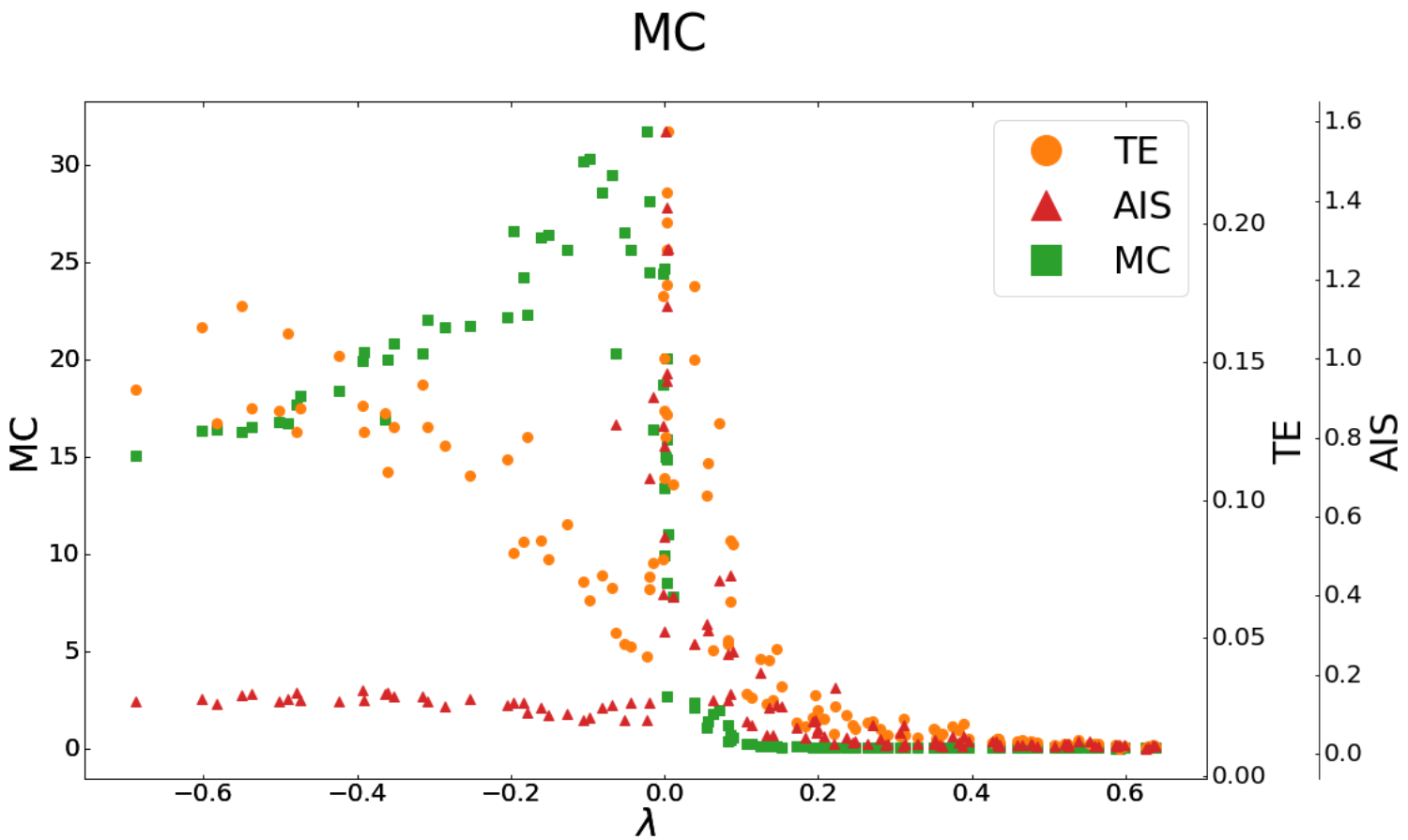
- fixed $\mathbf{w}^{\text{in}}$
- $w_{ij}^{\text{res}} \in \mathcal{N}(0; \sigma)$ such that $\log \sigma \in [-1.5; -0.25]$ in 26 steps 5 instances per value $\sigma$

**Information measures:**

- TE
  - Target - $k = 2$, $\tau_k = 1$
  - Source - $l = 1$, $\tau_l = 1$
- AIS - $k = 2$, $\tau_k = 1$
- 4-NN

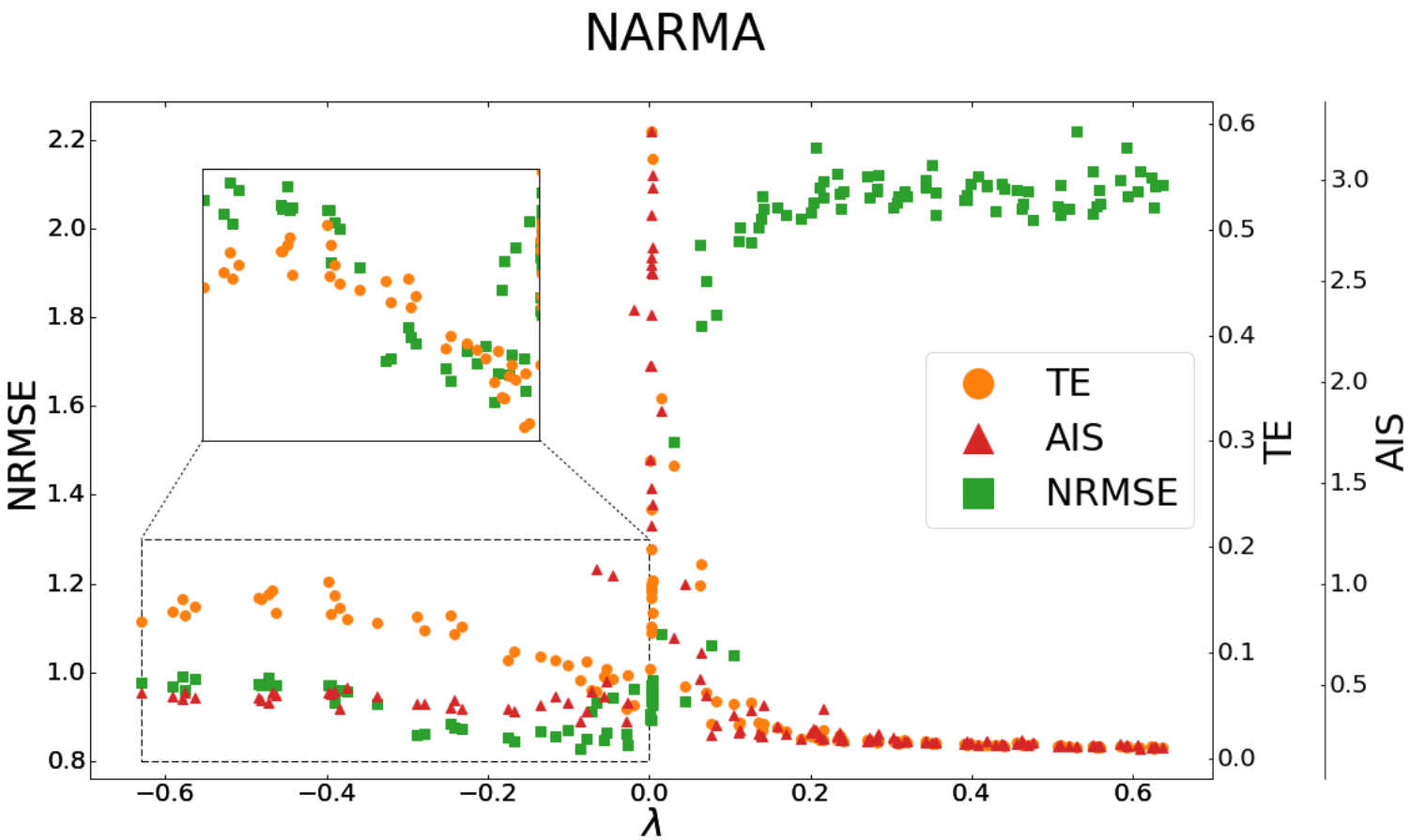
# Information-theoretical measures around criticality

## MC task



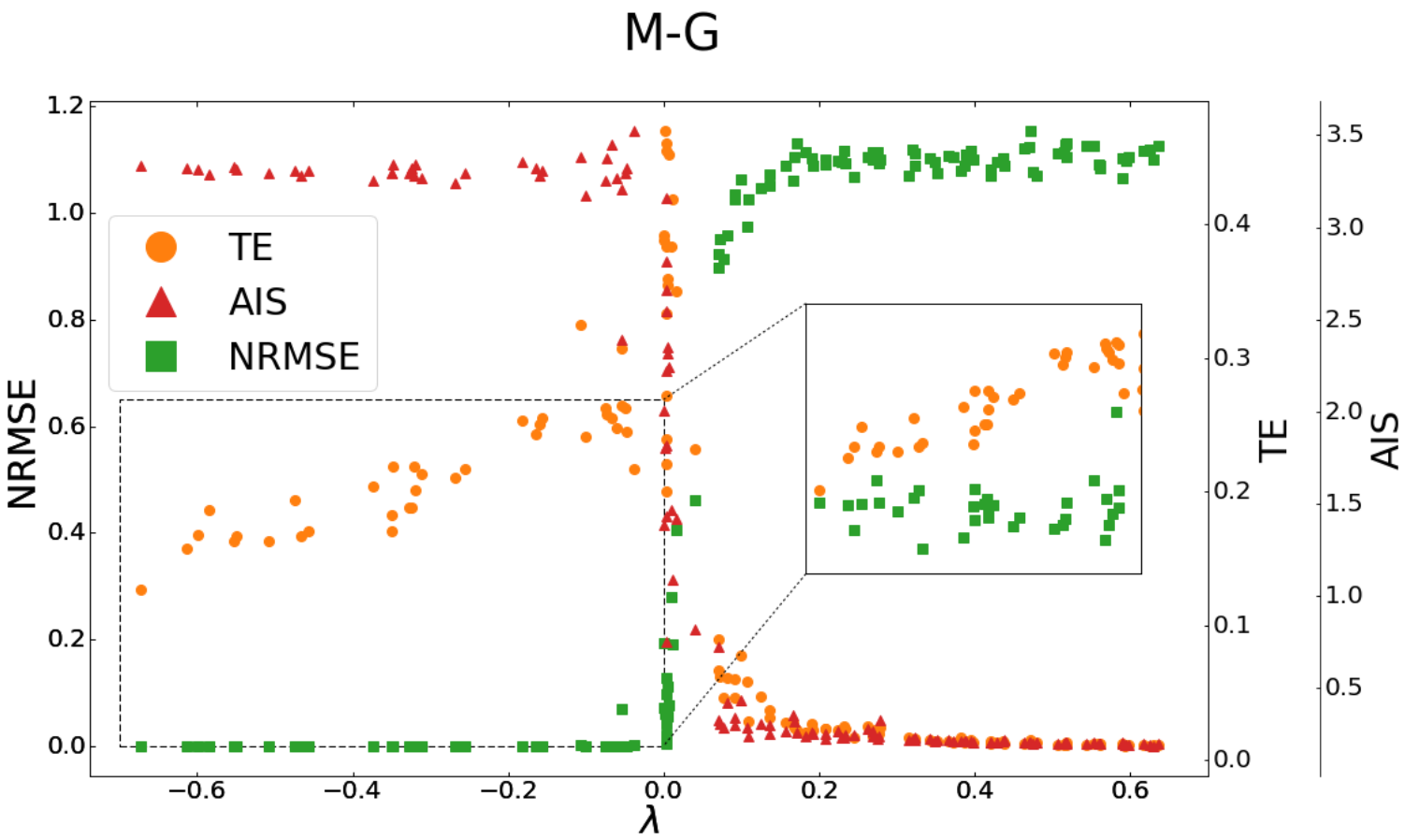
# Information-theoretical measures around criticality

## NARMA task



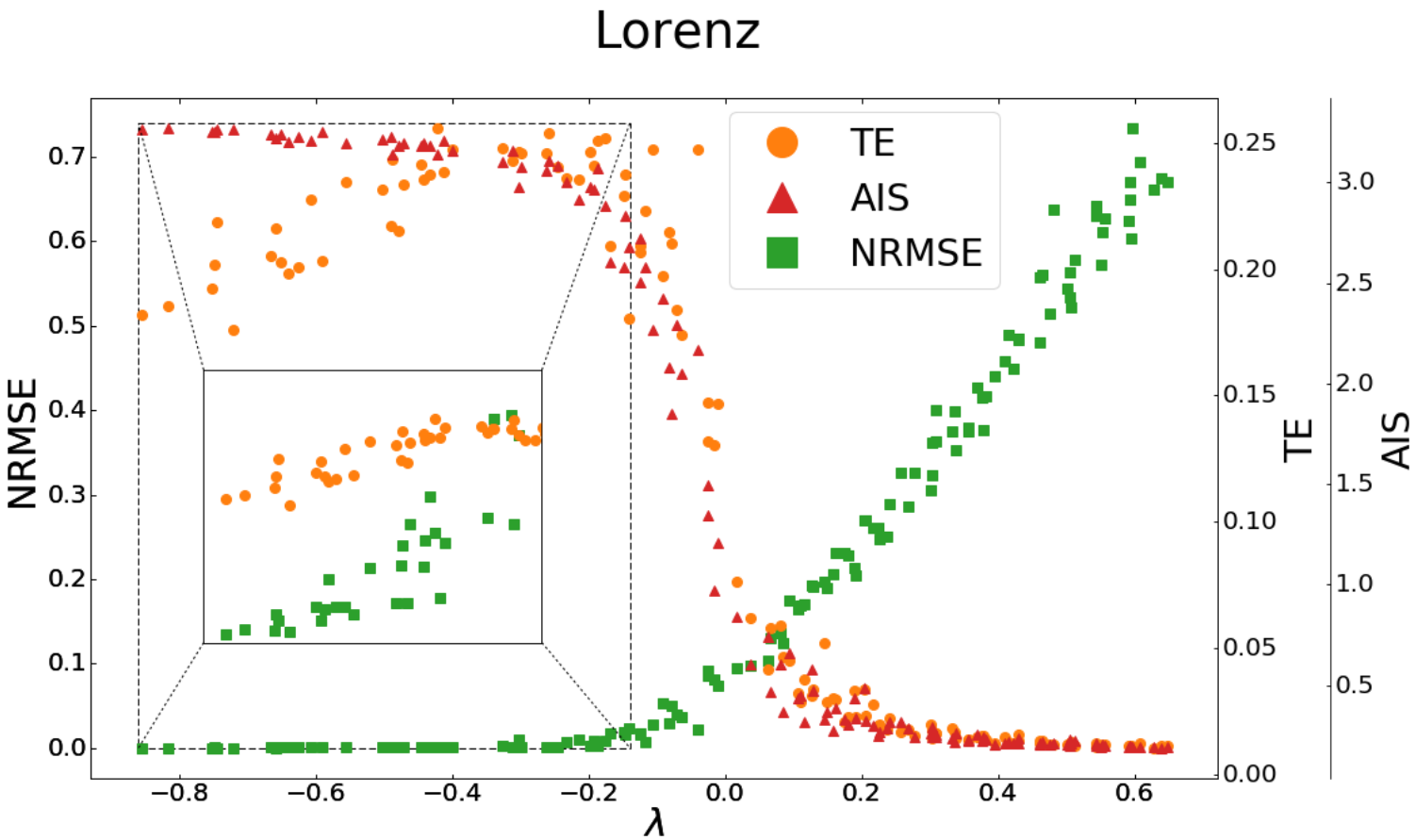
# Information-theoretical measures around criticality

## M–G task



# Information-theoretical measures around criticality

## Lorenz task



**Ragwitz-Kantz criterion:**

$$(k,\tau) = \underset{k\in\mathcal{Z},\tau\in\mathcal{Z}}{\operatorname{argmin}} \|\mathbf{x} - \widehat{\mathbf{x}}(k,\tau)\|$$

where

$$\mathbf{x} = (x_1,\ldots,x_n),\ \hat{\mathbf{x}} = (\widehat{x}_1,\ldots,\widehat{x}_n)$$

**Locally constant predictor:**

$$\widehat{x}_{t+1} = \frac{1}{\operatorname{card}(\mathcal{U}_t)} \sum_{\mathbf{x_l}\in\mathcal{U}_t} x_{l+1}$$

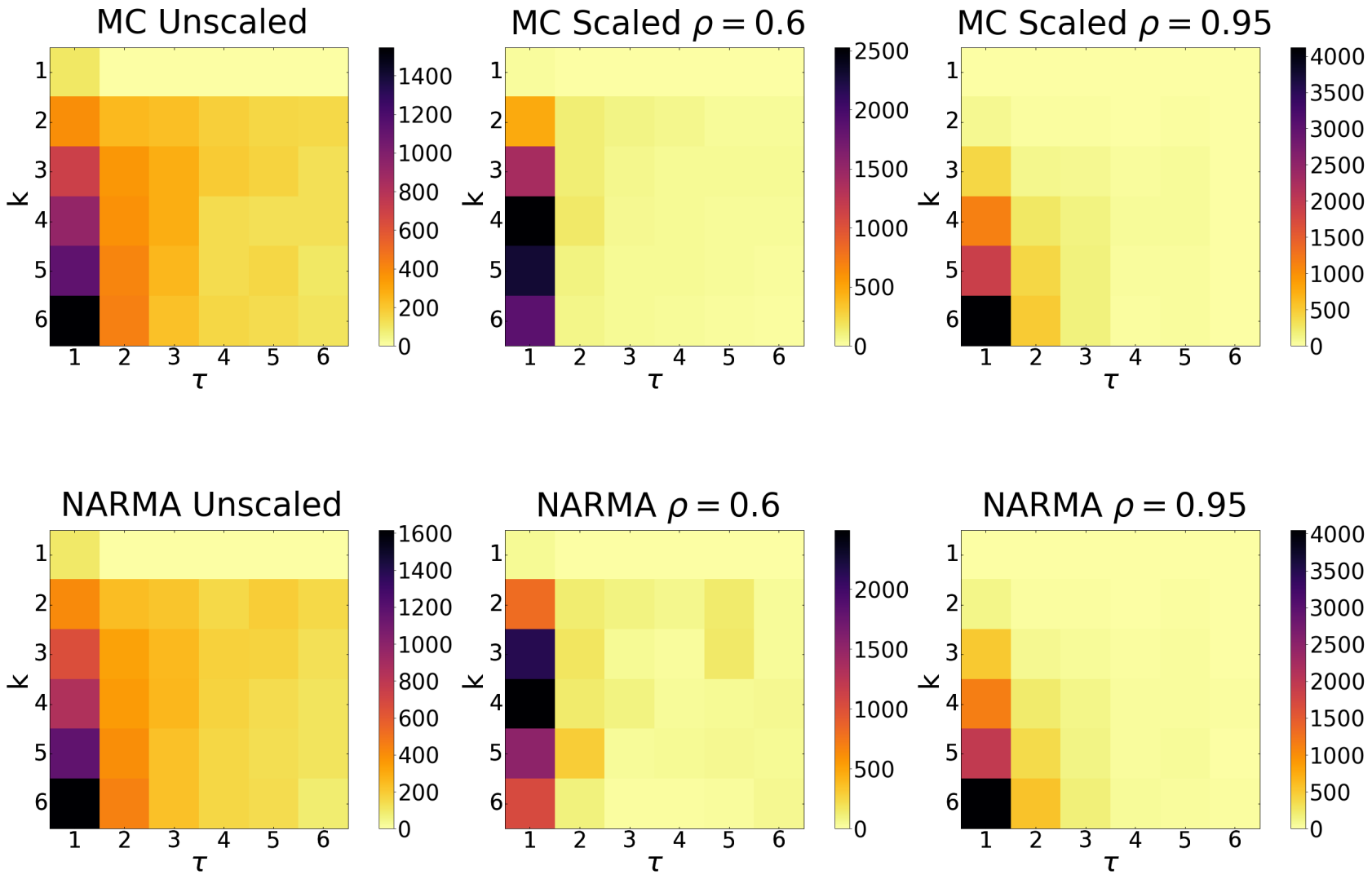
$$\mathcal{U}_t = \{\mathbf{x}_l : \|\mathbf{x}_l - \mathbf{x}_t\| \leq \epsilon\},\ \mathbf{x}_t = (x_t, x_{t-\tau_k}, x_{t-2\tau_k}, \ldots, , x_{t-(k-1)\tau})$$

# A closer look at information measures
## Reservoir neuron signal embedding: Experimental setup

- Ragwitz-Kantz search constrains:
  - $k = 1, \ldots, 6$
  - $\tau = 1, \ldots, 6$
- Reservoir spectral radius scalings:
  - Unstable: Unscaled
  - Stable: $\rho = 0.6$
  - Close to criticality: $\rho = 0.95$
- 100 instances per scaling and task

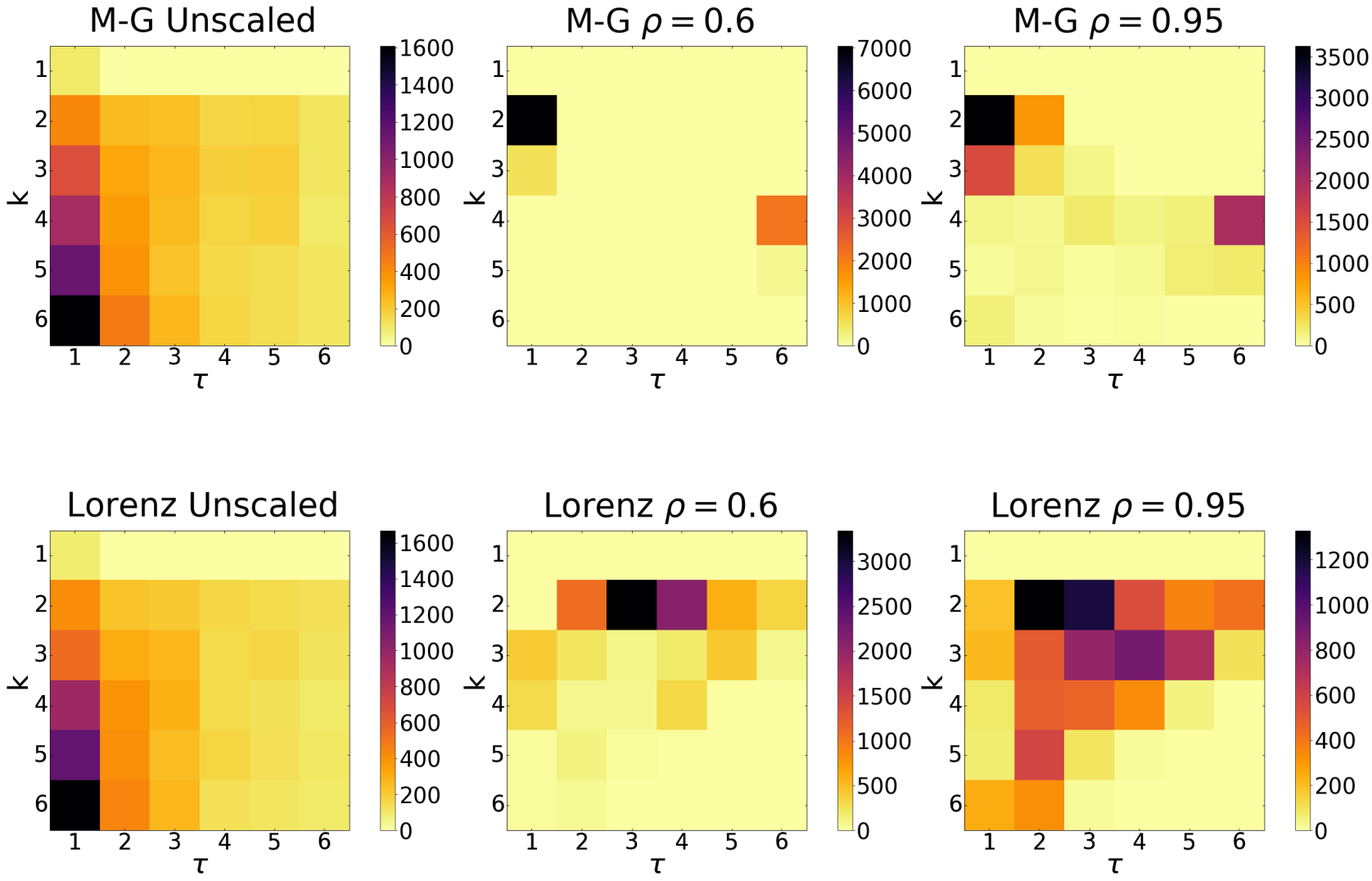
# A closer look at information measures

## Reservoir neuron signal embedding: MC and NARMA tasks



# A closer look at information measures

## Reservoir neuron signal embedding: M–G and Lorenz tasks



**Most frequent $(k, \tau)$ pairs for each task and scaling**

| Scaling \ Task | MC | | NARMA | | M–G | | Lorenz | |
|---|---|---|---|---|---|---|---|---|
| | $k$ | $\tau$ | $k$ | $\tau$ | $k$ | $\tau$ | $k$ | $\tau$ |
| init | 6 | 1 | 6 | 1 | 6 | 1 | 6 | 1 |
| $\rho = 0.6$ | 4 | 1 | 4 | 1 | 2 | 1 | 2 | 3 |
| $\rho = 0.95$ | 6 | 1 | 6 | 1 | 2 | 1 | 2 | 2 |

# A closer look at information measures
## Experimental setup

**ESN:**

- 1 instance for all tasks and scalings
  - $w_j^{in} \in \mathcal{U}(-0.1; 0.1), \forall j = 1, \ldots, N$
  - $w_{ij}^{res} \in \mathcal{N}(0; 0.5), \forall i, j = 1, \ldots, N$
- Reservoir spectral radius scalings
  - Unstable: Unscaled
  - Stable: $\rho = 0.6$
  - Close to criticality: $\rho = 0.95$

**Information measures:**

- TE
  - Target - determined by Ragwitz-Kantz criterion
  - Source - $l = 1$, $\tau_l = 1$
- AIS - determined by Ragwitz-Kantz criterion
- 4-NN

$$H_0\colon T_{Y\to X} = 0 \text{ vs. } H_1\colon T_{Y\to X} > 0$$

*p-value:*

$$P(T_{Y_{sur}\to X} \geq T_{Y\to X}) = \sum_{Y_{sur}\colon \widehat{T}_{Y_{sur}\to X} \geq \widehat{T}_{Y\to X}} P(Y_{sur})$$

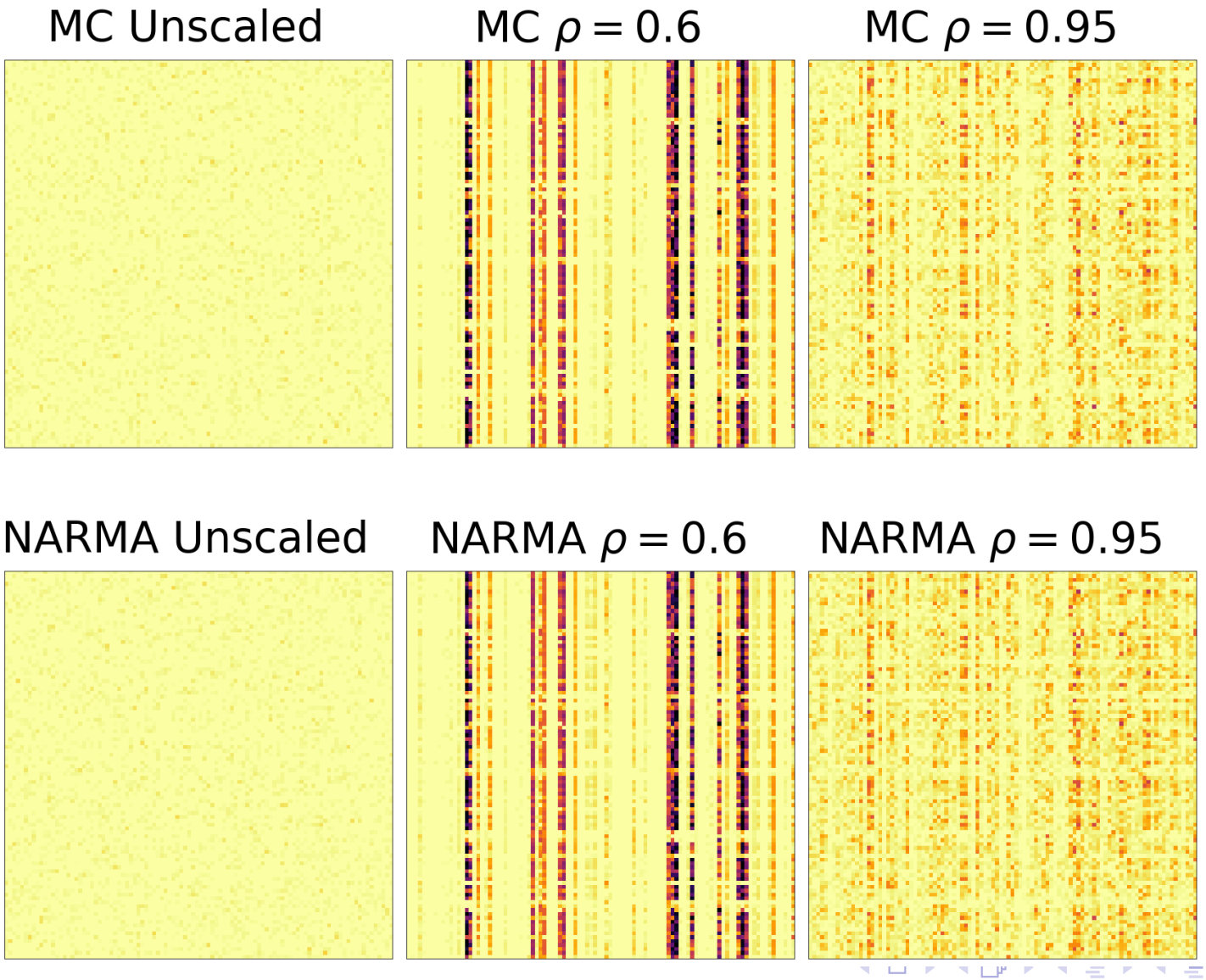
where $Y_{sur}$ is permutaion of

$$Y = (\mathbf{y}_n^{(k)}, \mathbf{y}_{n-1}^{(k)}, \ldots, \mathbf{y}_1^{(k)}).$$

- p-value was computed from 100 $\widehat{T}_{Y_{sur}}$
- $\widehat{T}_{Y\to X} \leftarrow 0$ if $P(T_{Y_{sur}\to X} \geq T_{Y\to X}) > 0.05$

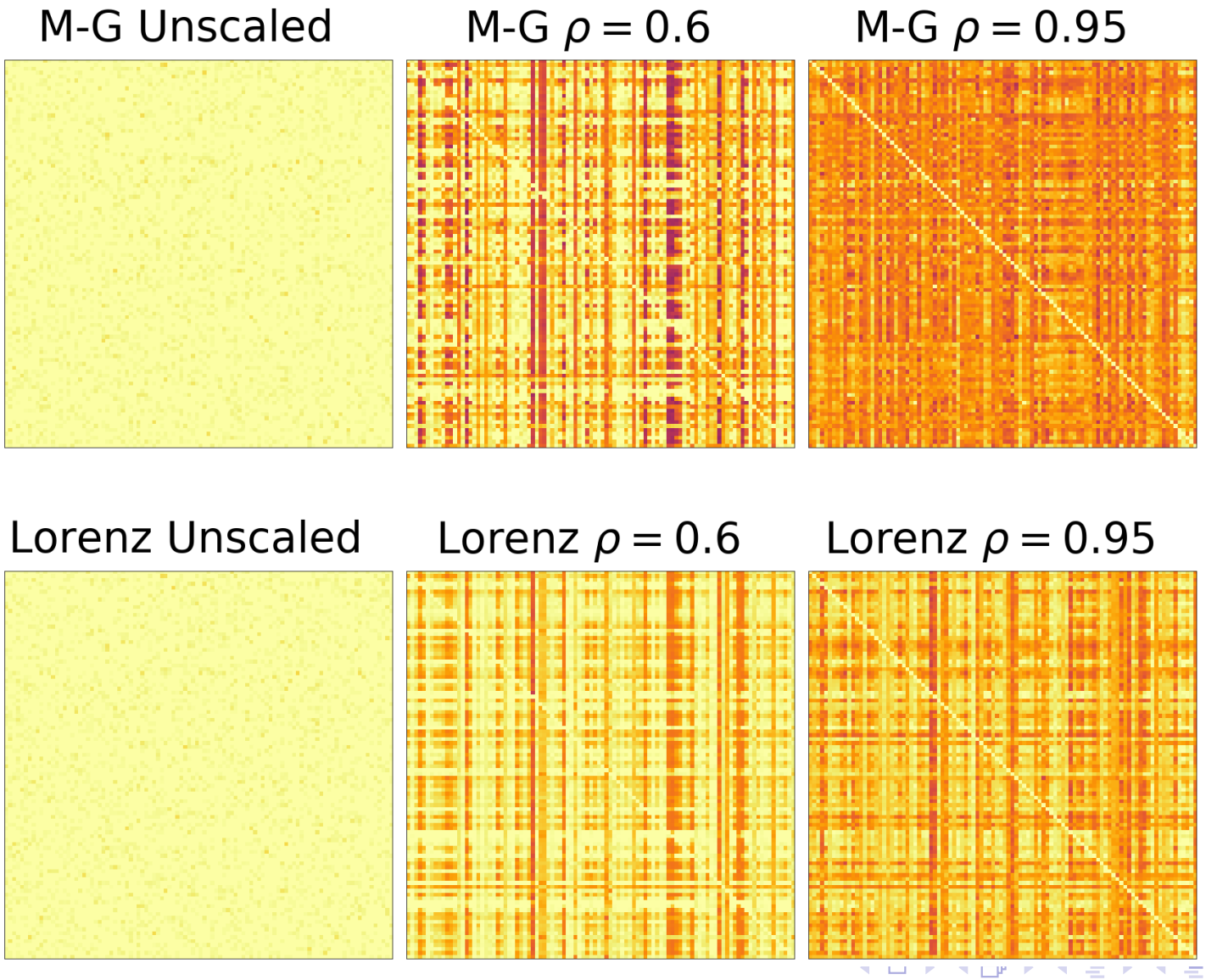
# A closer look at information measures

## TE matrices: MC and NARMA tasks



# A closer look at information measures

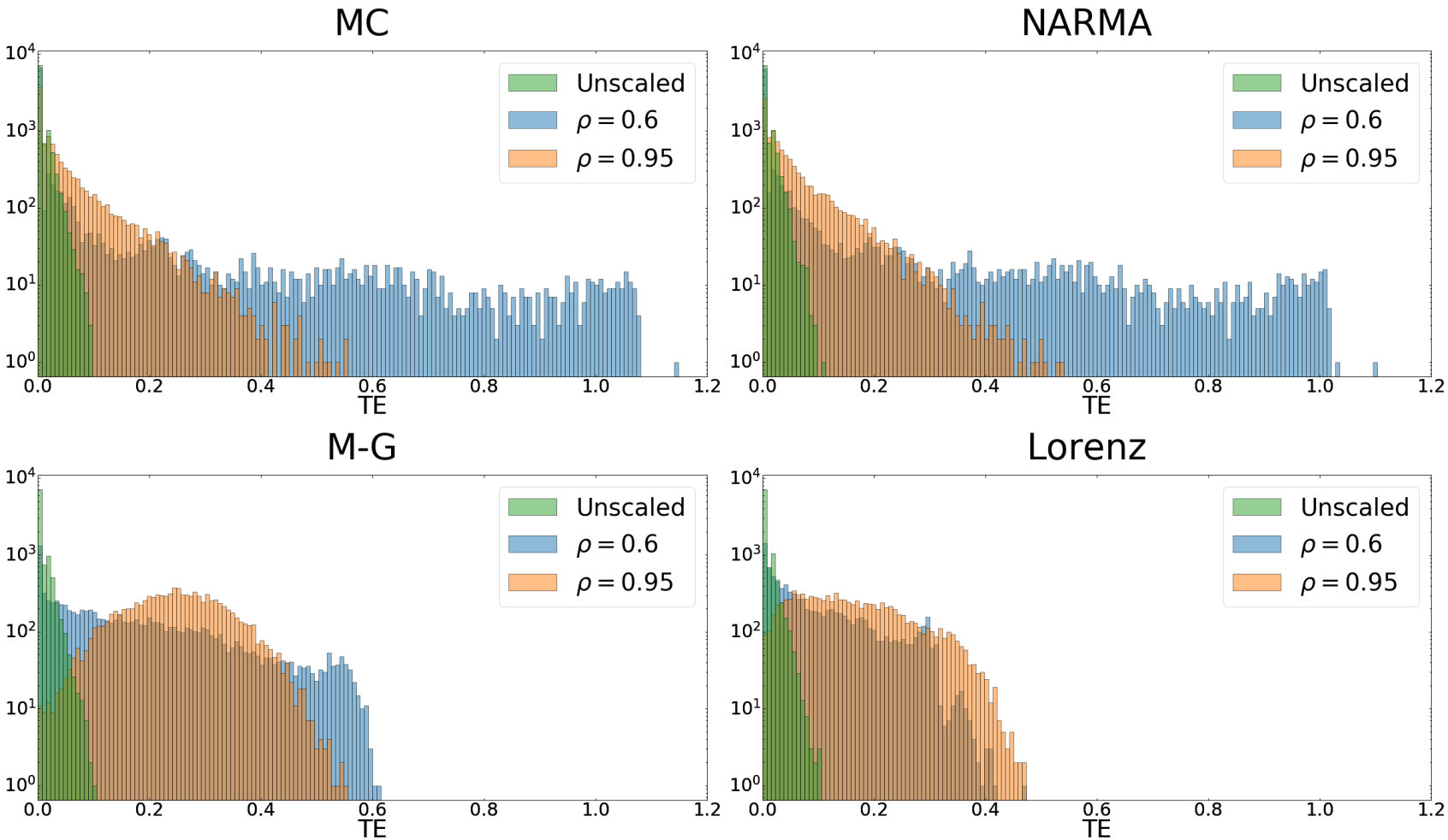
## TE matrices: M–G and Lorenz tasks



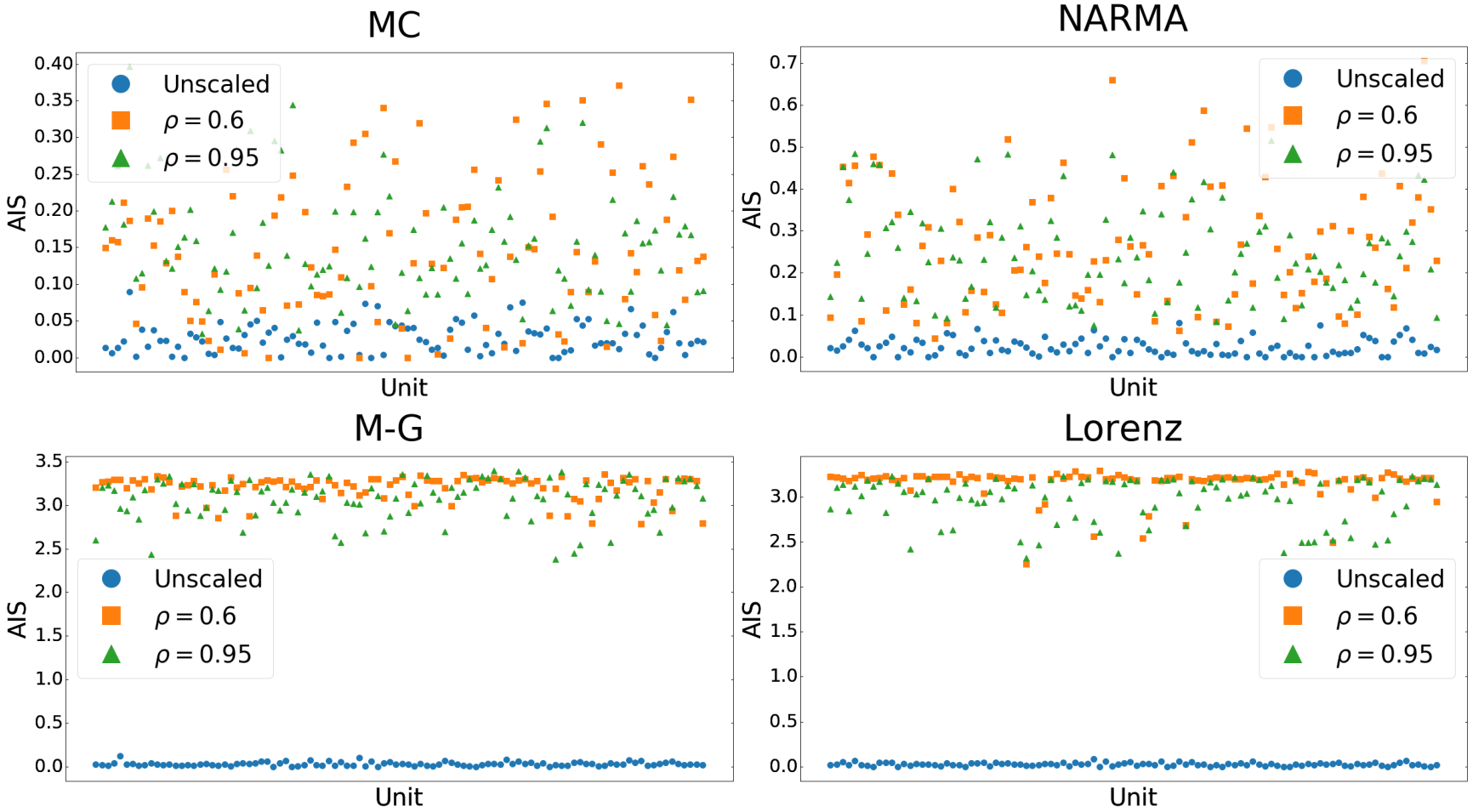
$$\left|\overline{H}(\mathrm{TE}_{\mathrm{RES}}^{(k,l)})\right| = \widehat{D_{\mathrm{KL}}}(\mathrm{TE}_{\mathrm{RES}}^{(k,l)} \,\|\, \mathcal{U}(0; \max \mathrm{TE}_{\mathrm{RES}}^{(k,l)}))$$
$$= \ln(\max \mathrm{TE}_{\mathrm{RES}}^{(k,l)}) - \widehat{H_{\mathrm{KL}}}(\mathrm{TE}_{\mathrm{RES}}^{(k,l)})$$

where $\ln(\max \mathrm{TE}_{\mathrm{RES}}^{(k,l)})$ is the exact differential entropy of the uniform distribution with the support in $[0; \max \mathrm{TE}_{\mathrm{RES}}^{(k,l)}]$, $\widehat{H_{\mathrm{KL}}}(\cdot)$ is the Kozachenko-Leonenko differential entropy estimator and $\mathrm{TE}_{\mathrm{RES}}^{(k,l)}$ is the distribution of the estimates of transfer entropies in the reservoir.

# A closer look at information measures
## TE distribution



# A closer look at information measures
## Active Information Storage



# A closer look at information measures
## Quantification of used measures

| **MC task** | Unscaled | $\rho = 0.6$ | $\rho = 0.95$ |
|---|---|---|---|
| MC | 0.06 | 17.8 | 32.8 |
| Average TE | 0.007 | 0.091 | 0.047 |
| Average AIS | 0.027 | 0.146 | 0.151 |
| Rel. entr. of TE | 0.981 | 0.629 | 1.147 |
| LE | 0.53 | $-0.52$ | $-0.06$ |

| **NARMA** | Unscaled | $\rho = 0.6$ | $\rho = 0.95$ |
|---|---|---|---|
| NRMSE | 2.143 | 0.98 | 0.83 |
| Average TE | 0.007 | 0.087 | 0.052 |
| Average AIS | 0.023 | 0.267 | 0.247 |
| Rel. entr. of TE | 1.086 | 0.783 | 1.174 |
| LE | 0.52 | $-0.53$ | $-0.062$ |

| **Mackey–Glass** | Unscaled | $\rho = 0.6$ | $\rho = 0.95$ |
|---|---|---|---|
| NRMSE | 1.13 | 0.00025 | 0.00026 |
| Average TE | 0.007 | 0.154 | 0.248 |
| Average AIS | 0.029 | 3.204 | 3.073 |
| Rel. entr. of TE | 0.266 | 0.386 | 0.587 |
| LE | 0.53 | $-0.52$ | $-0.062$ |

| **Lorenz** | Unscaled | $\rho = 0.6$ | $\rho = 0.95$ |
|---|---|---|---|
| NRMSE | 0.54 | 0.00012 | 0.0013 |
| Average TE | 0.007 | 0.087 | 0.157 |
| Average AIS | 0.025 | 3.157 | 2.971 |
| Rel. entr. of TE | 1.018 | 0.642 | 0.231 |
| LE | 0.52 | $-0.74$ | $-0.28$ |

# Conclusion

- Two complexity classes in tested tasks
  - stochastic: MC/NARMA - benefit from criticality
  - deterministic: M–G/Lorenz - do not benefit from criticality
- Information measures are maximized at phase transition from stable to unstable regime
- Between complexity classes: increase in performance is associated with increase in global TE and AIS, and decrease of relative entropy of TE
- Within tasks: in stable regions there seems to be an inverse relationship between performance and global TE.
- Higher information flow does not mean better performance.
- Task specific

# THANK YOU