Long short-term memory (LSTM) networks for the analysis of time series

(CIV6540 - Probabilistic Machine Learning for Civil Engineers)

Professor: James-A. Goulet

Département des génies civil, géologique et des mines
Polytechnique Montréal

Chapter 10 – Goodfellow et al. (2016)
Deep Learning, MIT Press

Olah (2015)
Understanding LSTM Networks, (blog)

Module #7c Outline

Introduction
Feedforward NN (review)
Recurrent NN
LSTM
Example

Topics organization

Background
- 1 Revision probability & linear algebra
- 2 Probability distributions

Machine Learning Basics
- 0 Introduction
- 3 Bayesian Estimation $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$
- 4 MCMC sampling & Newton

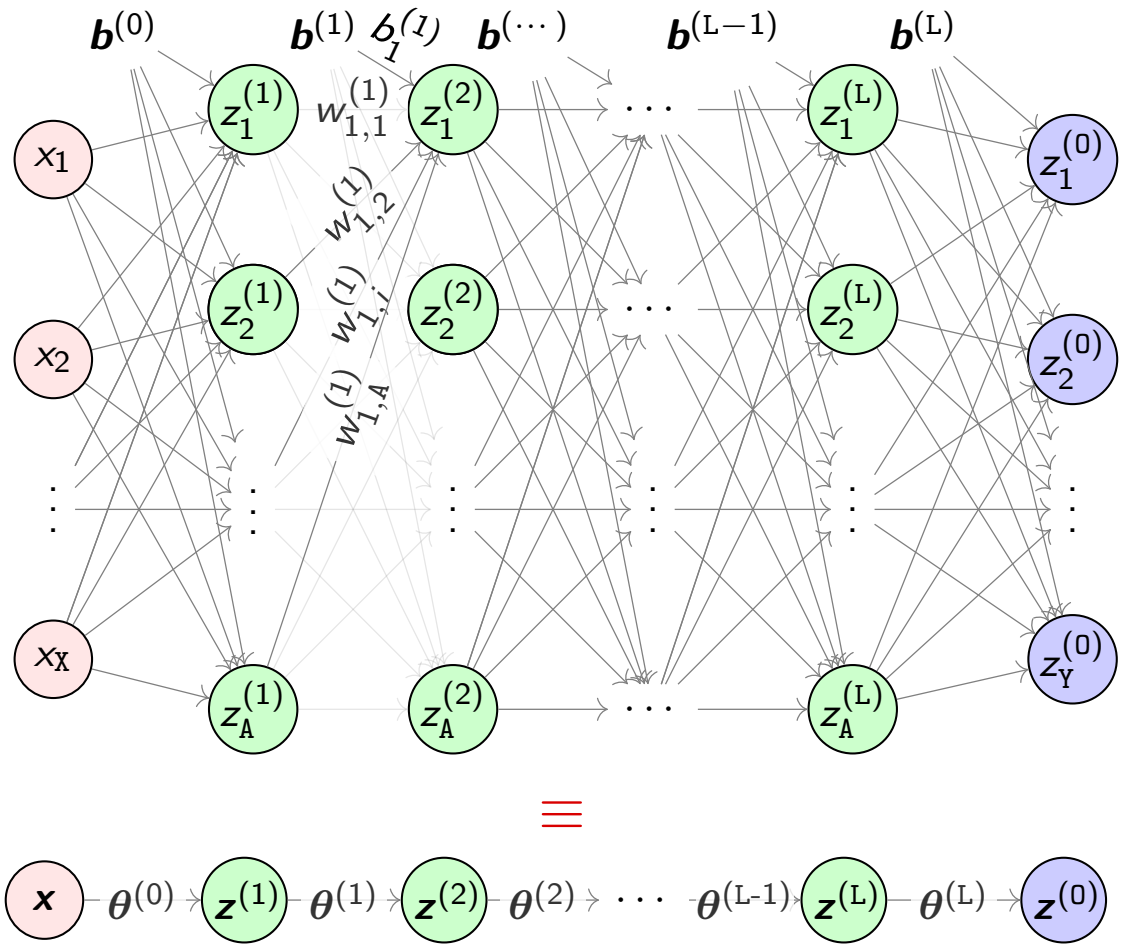
Supervised learning
- 5 Regression
- 6 Classification
- 7 LSTM networks for time series

Unsupervised learning
- 7 State-space model for time-series

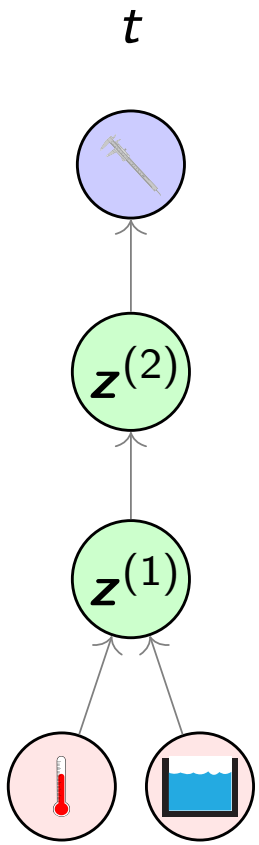
Decision Making & RL
- 8 Decision Theory
- 9 AI & Sequential decision problems

Feedforward NN Representation

- $\mathbf{x}$: Input layer
- $\mathbf{a}^{(l)}$: Activation units
- $\mathbf{z}^{(l)}$: Hidden units
- $\mathbf{z}^{(0)}$: Output layer
- $\boldsymbol{\theta}$: parameters

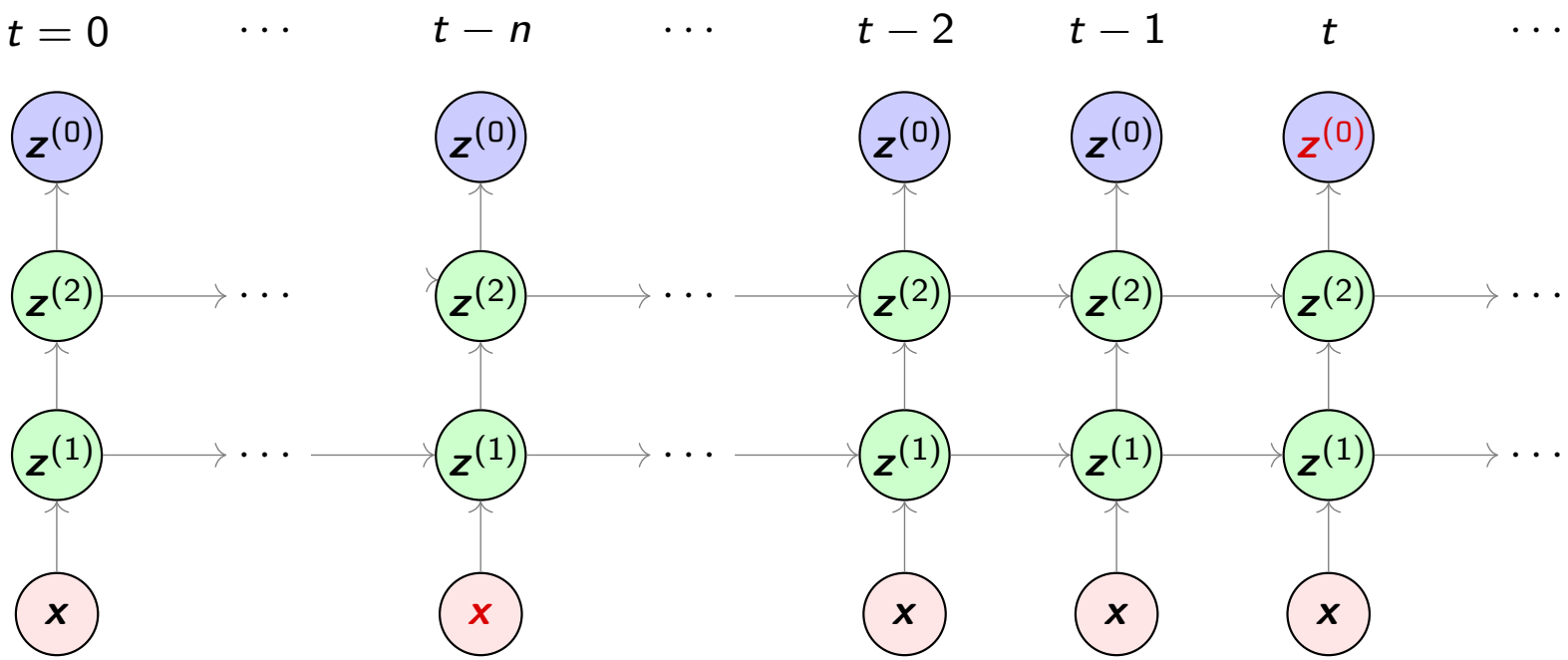


Example - Modelling the displacement of a dam



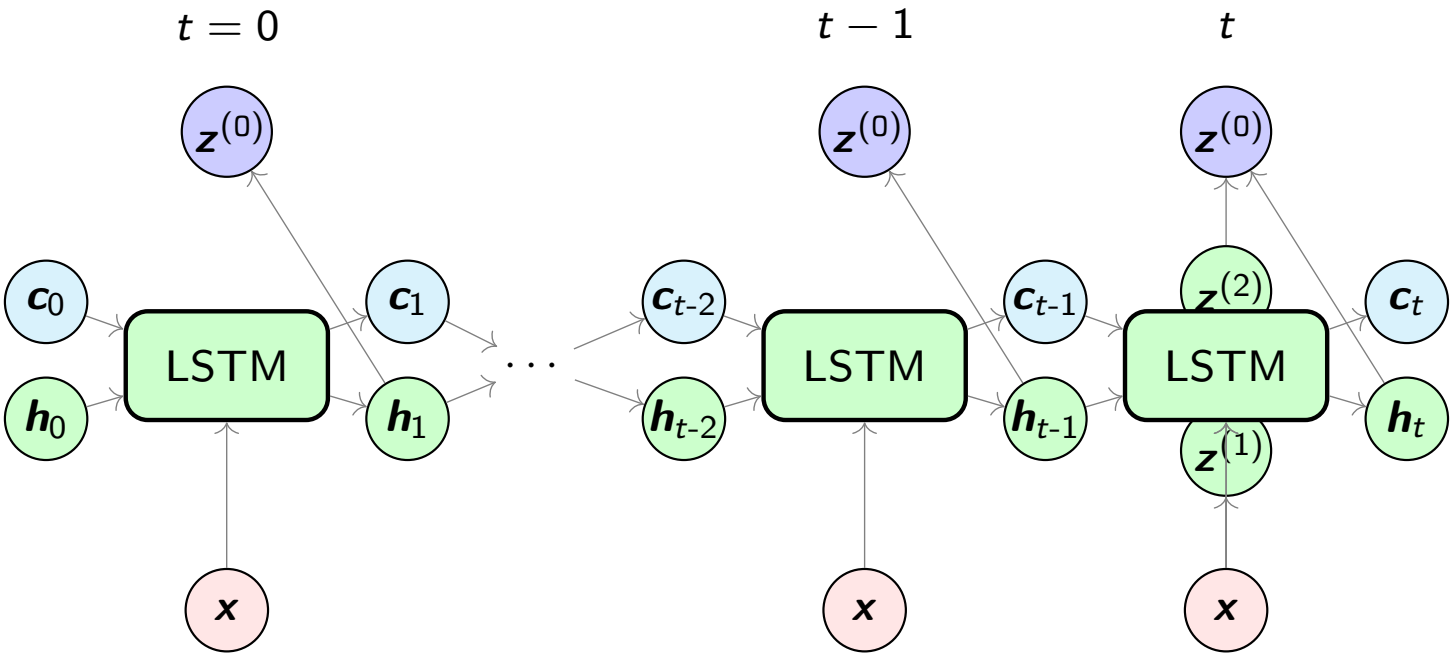
[Hydro-Quebec]

Recurrent neural networks



Vanishing gradient for long-term dependencies: $\frac{\partial z_t^{(0)}}{\partial x_{t-n}} \to 0$

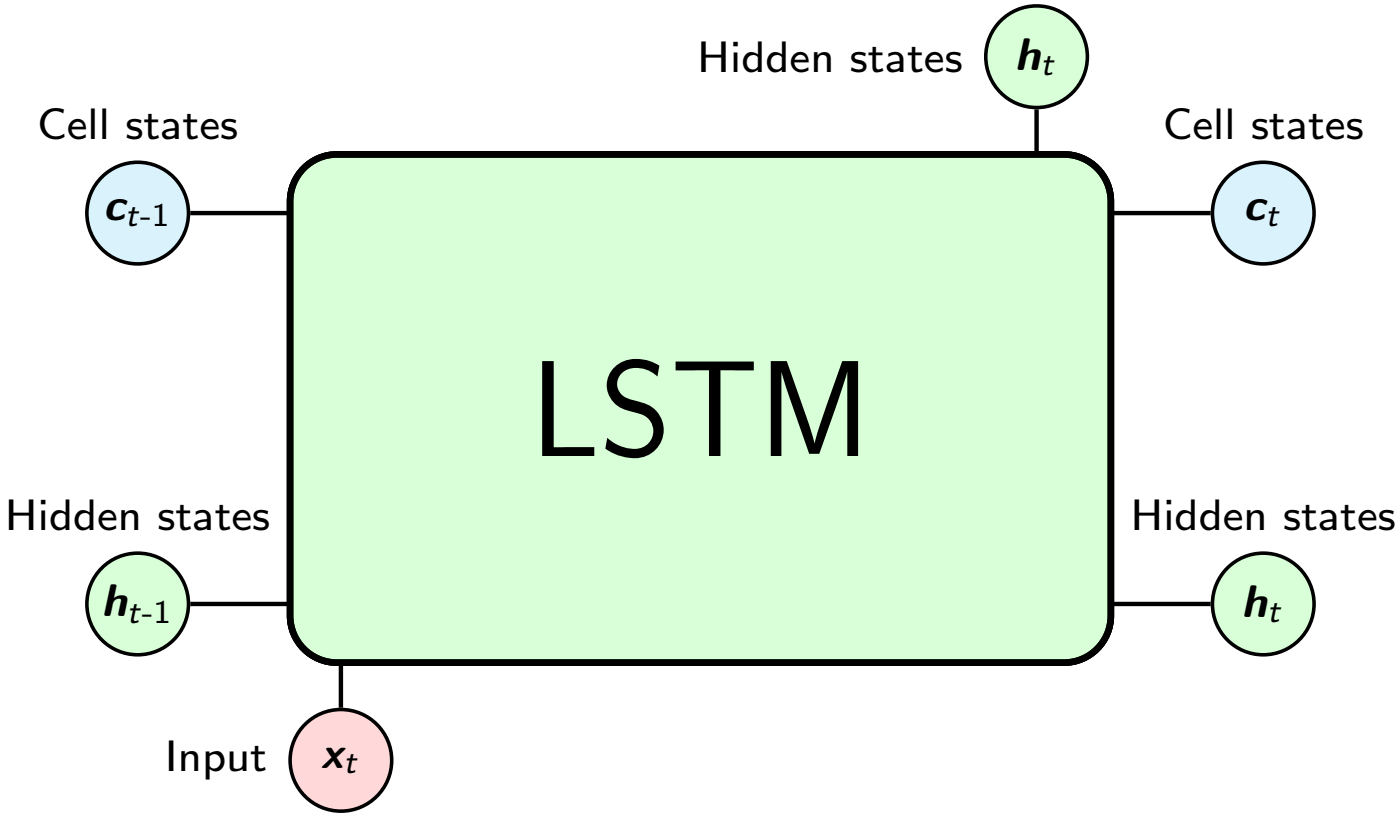
LSTM network



The key behind the LSTM is the cell state (memory)

$$\boldsymbol{c}_t = [c_1, c_2, \cdots, c_C]_t^\intercal \in \mathbb{R}^C$$

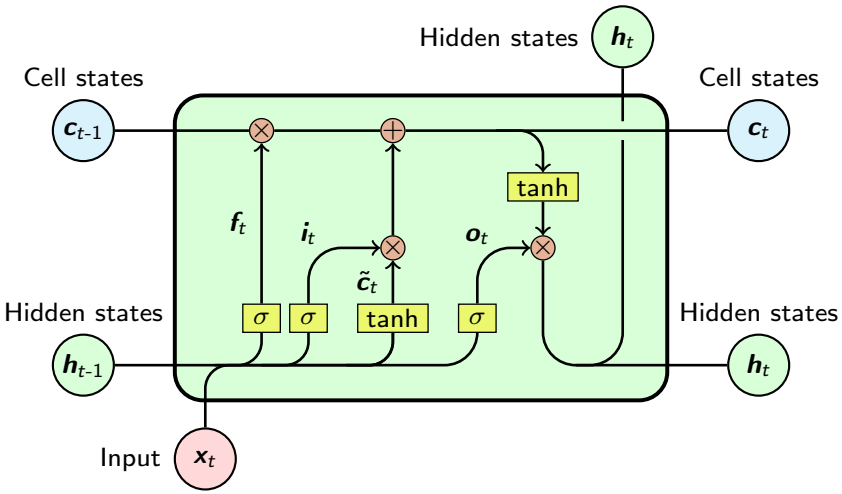
The LSTM cell

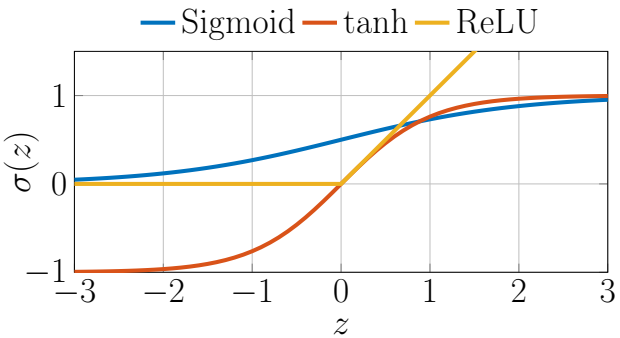


LSTM gates: forget irrelevant past information from c_{t-1},
write new relevant information in c_t,
and **rely on a subset** of c_t to make predictions h_t

[Adapted from J. Leon, Beerware]

The forget gate

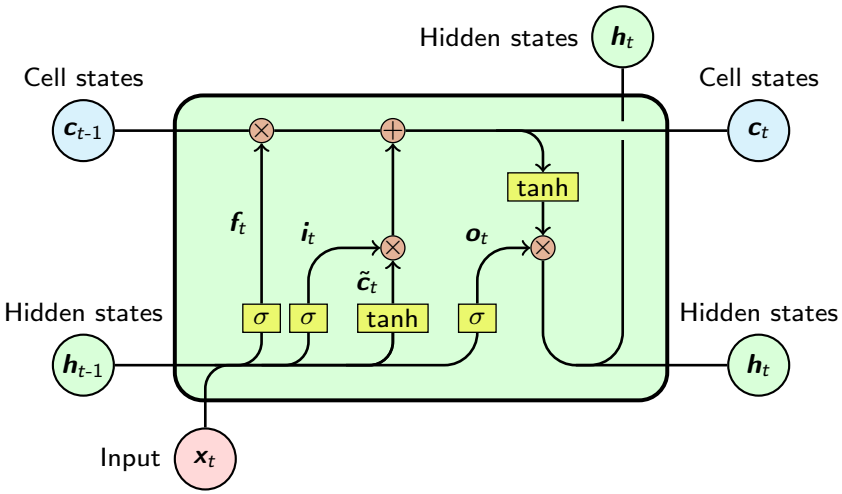




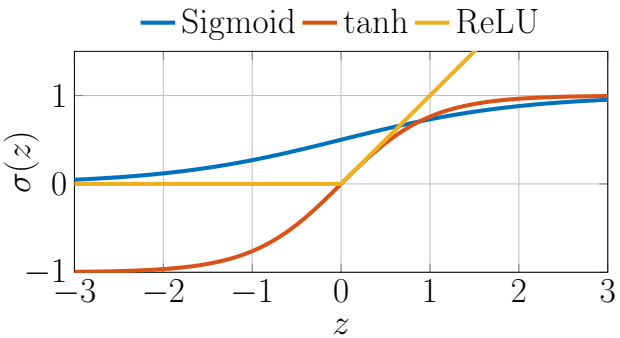
$$\boldsymbol{c}_{t-1} = [c_1, c_2, \cdots, c_C]^\intercal_{t-1} \in \mathbb{R}^C$$
$$\boldsymbol{f}_t = [f_1, f_2, \cdots, f_C]^\intercal_t \in (0,1)^C$$
$$\boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} = [c_1, 0, \cdots, c_C]^\intercal_t$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \cdot [\boldsymbol{h}_{t-1}; \boldsymbol{x}_t] + \boldsymbol{b}_f)$$

The input and candidate gates

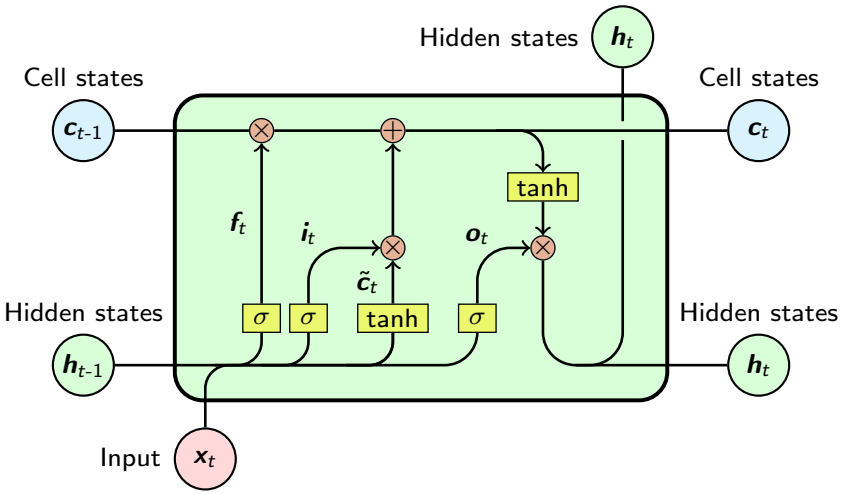


$$\tilde{\boldsymbol{c}}_t = [\tilde{c}_1, \tilde{c}_2, \cdots, \tilde{c}_C]_t^T \in (-1,1)^C$$
$$\boldsymbol{i}_t = [i_1, i_2, \cdots, i_C]_t^T \in (0,1)^C$$
$$\boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t = [\tilde{c}_1, 0, \cdots, \tilde{c}_C]_t^T$$

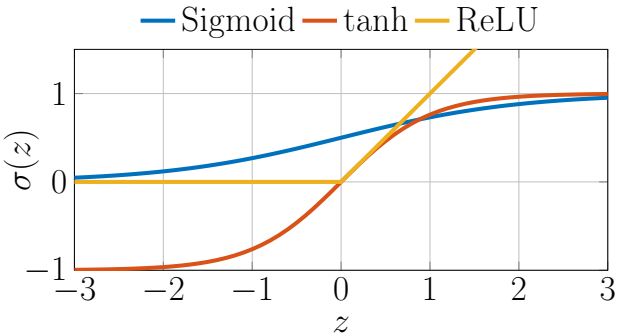


$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \cdot [\boldsymbol{h}_{t-1}; \boldsymbol{x}_t] + \boldsymbol{b}_i)$$
$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c \cdot [\boldsymbol{h}_{t-1}; \boldsymbol{x}_t] + \boldsymbol{b}_c)$$
$$\boldsymbol{c}_t = \underbrace{\boldsymbol{f}_t \odot \boldsymbol{c}_{t-1}}_{\text{forget}} + \underbrace{\boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t}_{\text{write}}$$

The output gate



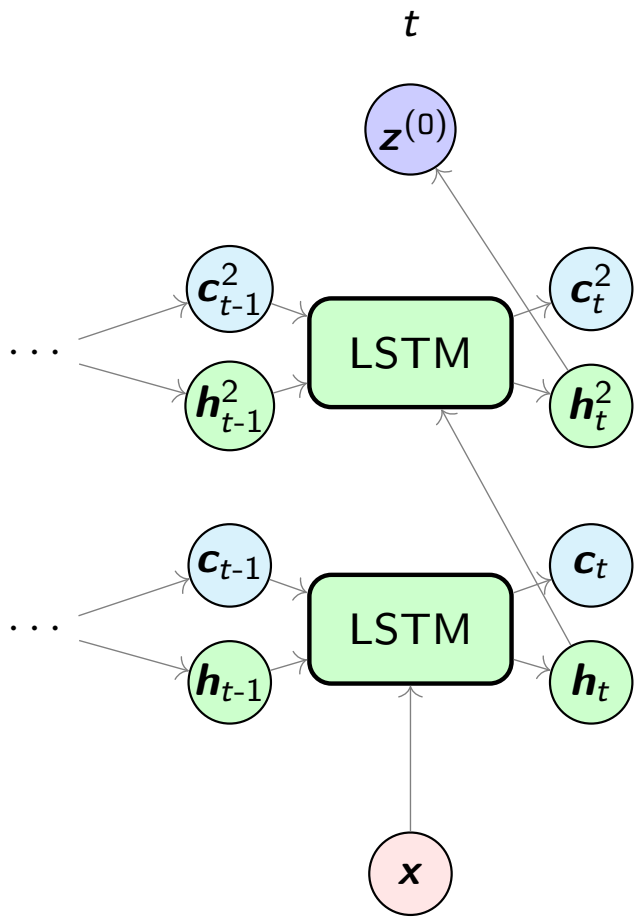
$$\boldsymbol{c}_t = [c_1, c_2, \cdots, c_C]_t^\intercal \in \mathbb{R}^C$$
$$\boldsymbol{o}_t = [o_1, o_2, \cdots, o_C]_t^\intercal \in (0,1)^C$$
$$\boldsymbol{o}_t \odot \boldsymbol{c}_t = [c_1, 0, \cdots, c_C]_t^\intercal$$



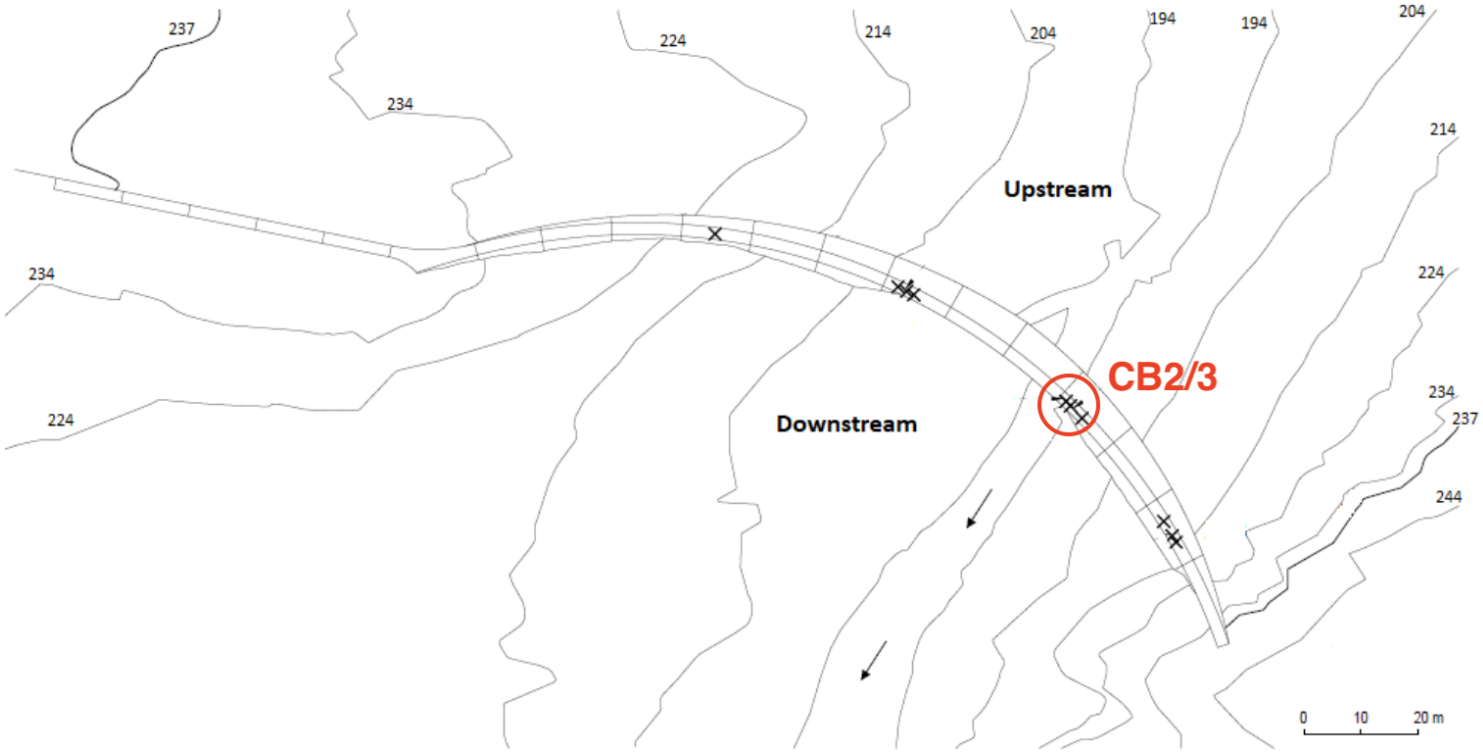
$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \cdot [\boldsymbol{h}_{t-1}; \boldsymbol{x}_t] + \boldsymbol{b}_o)$$
$$\boldsymbol{h}_t = \underbrace{\boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)}$$

Attention on a subset of the cell state

Multi-layers LSTM network

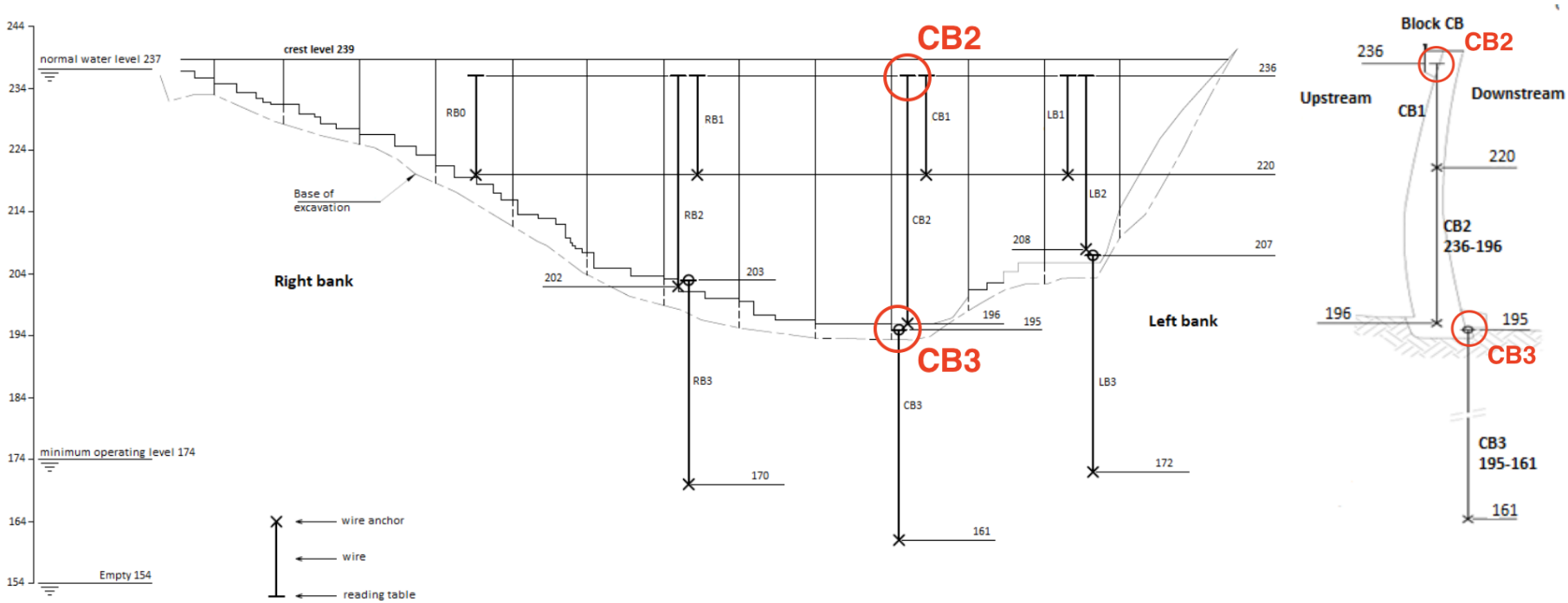


Dam structural health monitoring



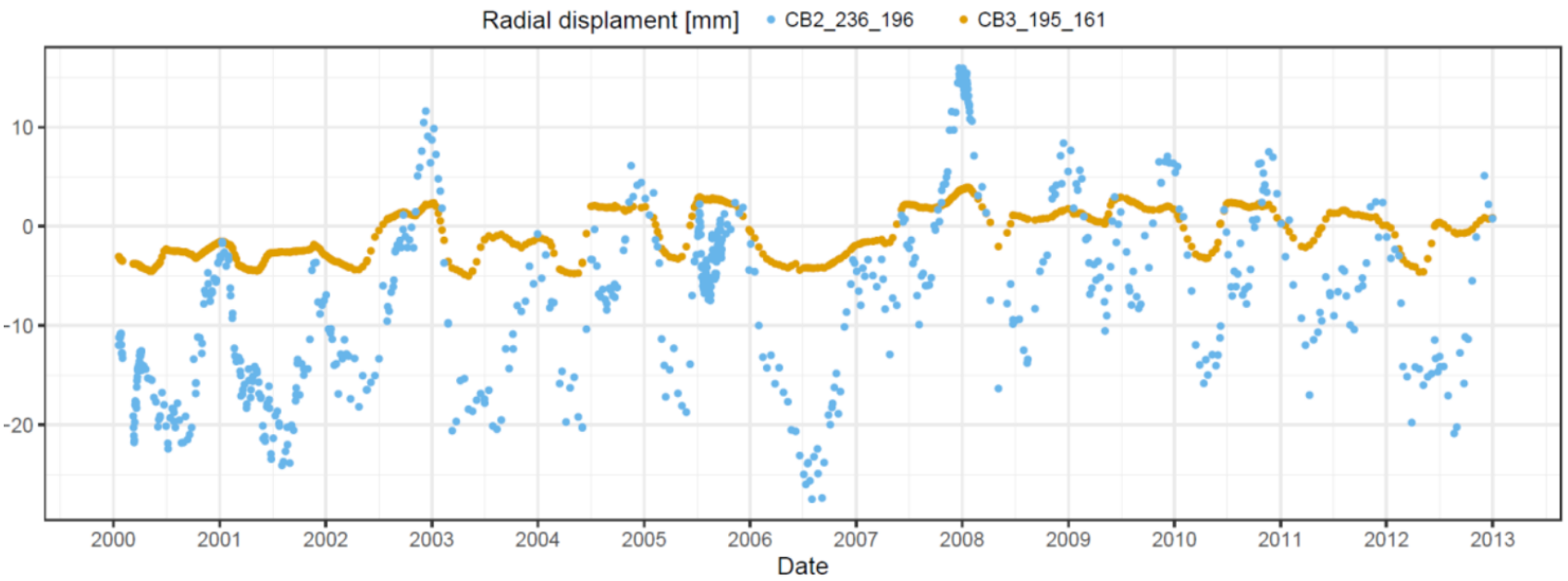
[ICOLD-BW2022]

Dam structural health monitoring



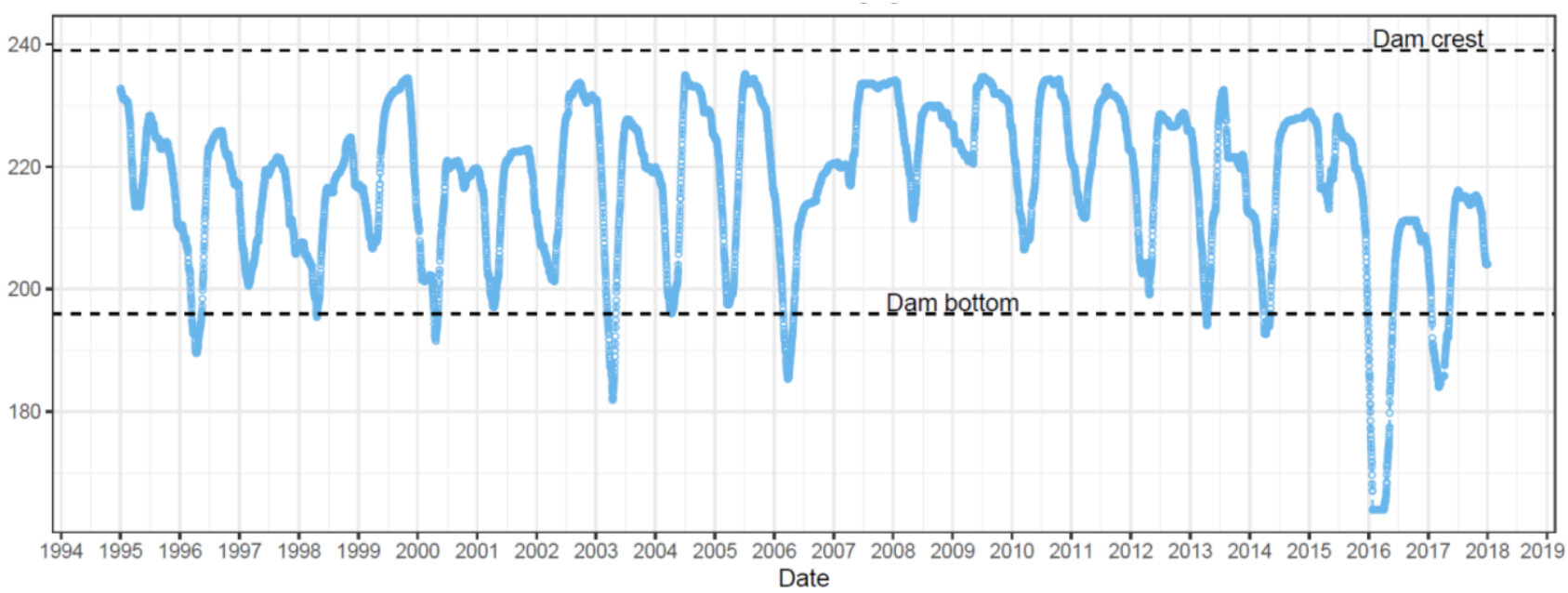
[ICOLD-BW2022]

Data – Pendulum ($\sim$ 1 datapoint /1.5 week)



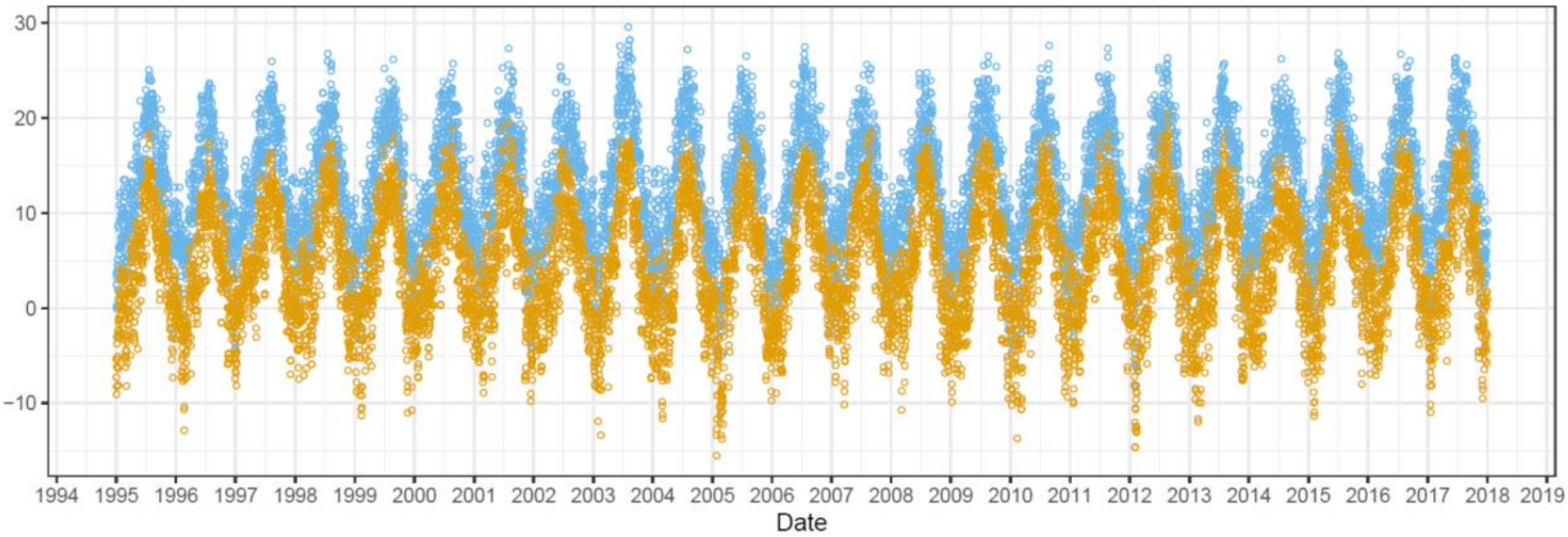
[ICOLD-BW2022]

Data – Reservoir level



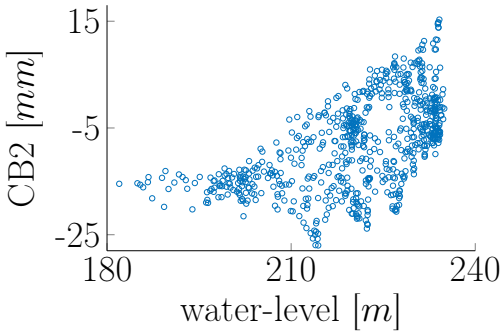
[ICOLD-BW2022]

Data – Temperature

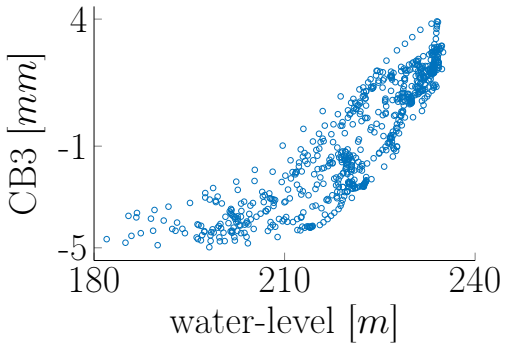


[ICOLD-BW2022]

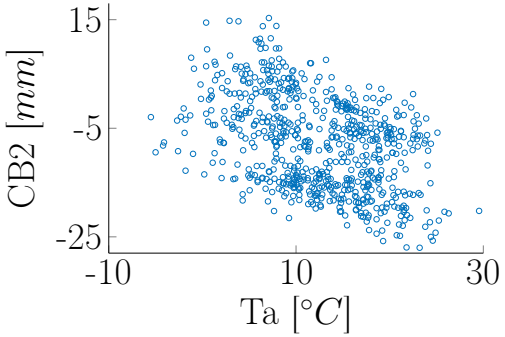
Relationships between datasets



Water level v.s. CB2

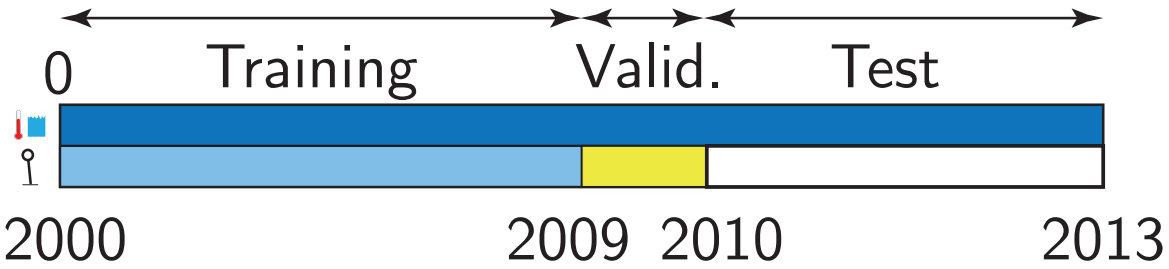


Water level v.s. CB3

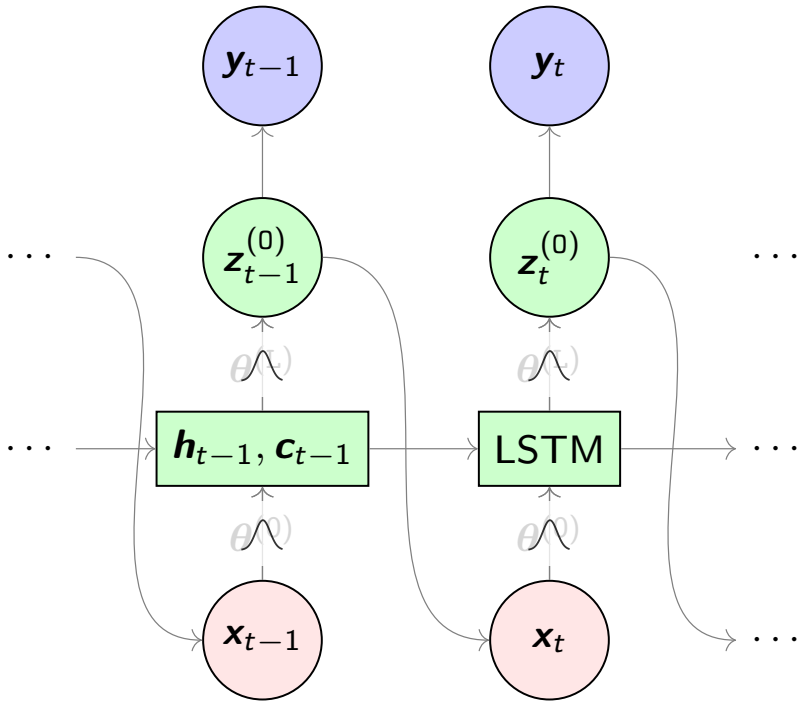


Temp. v.s. CB2 & CB3

Compare the predictive capacity of LSTM and BDLM



LSTM architecture with teacher forcing



We use **TAGI** instead of backprop to train the model...

Two sources of uncertainty:

- Parameters ($\boldsymbol{\theta}$)

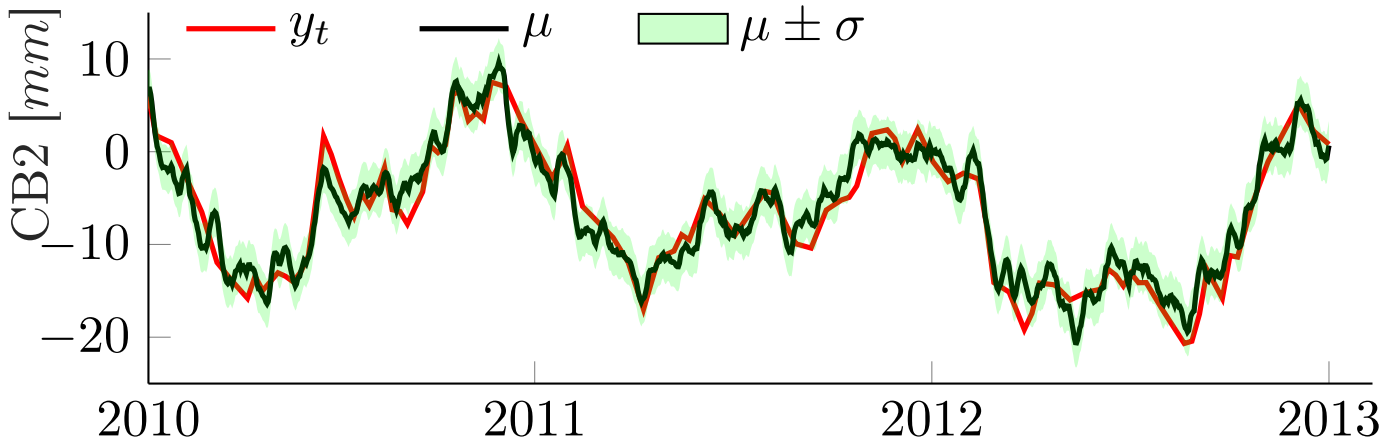
  $$\boldsymbol{\theta} = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$$

- Prediction error (w)

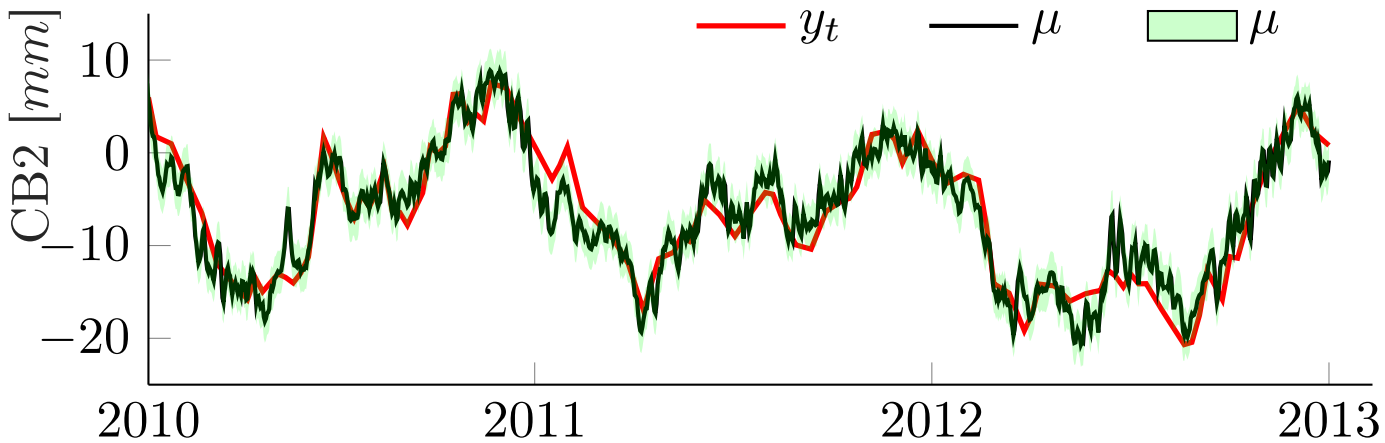
  $$y = z^{(0)} + w$$

Comparison of LSTM & BDLM – CB2

BDLM
($LL = -63.9$)

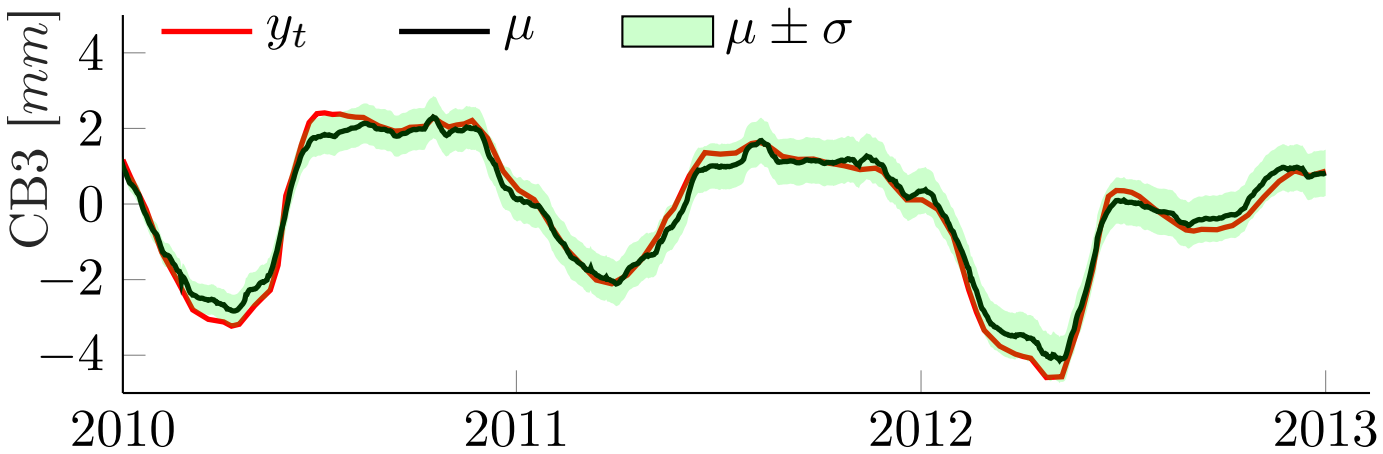


LSTM
($LL = -65.9$)

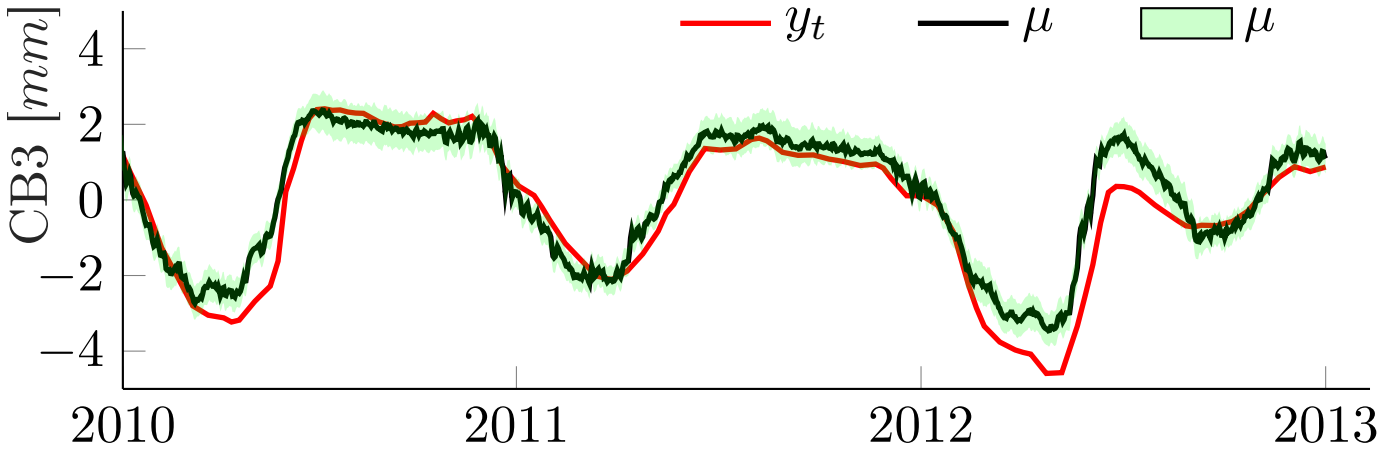


Comparison of LSTM & BDLM – CB3

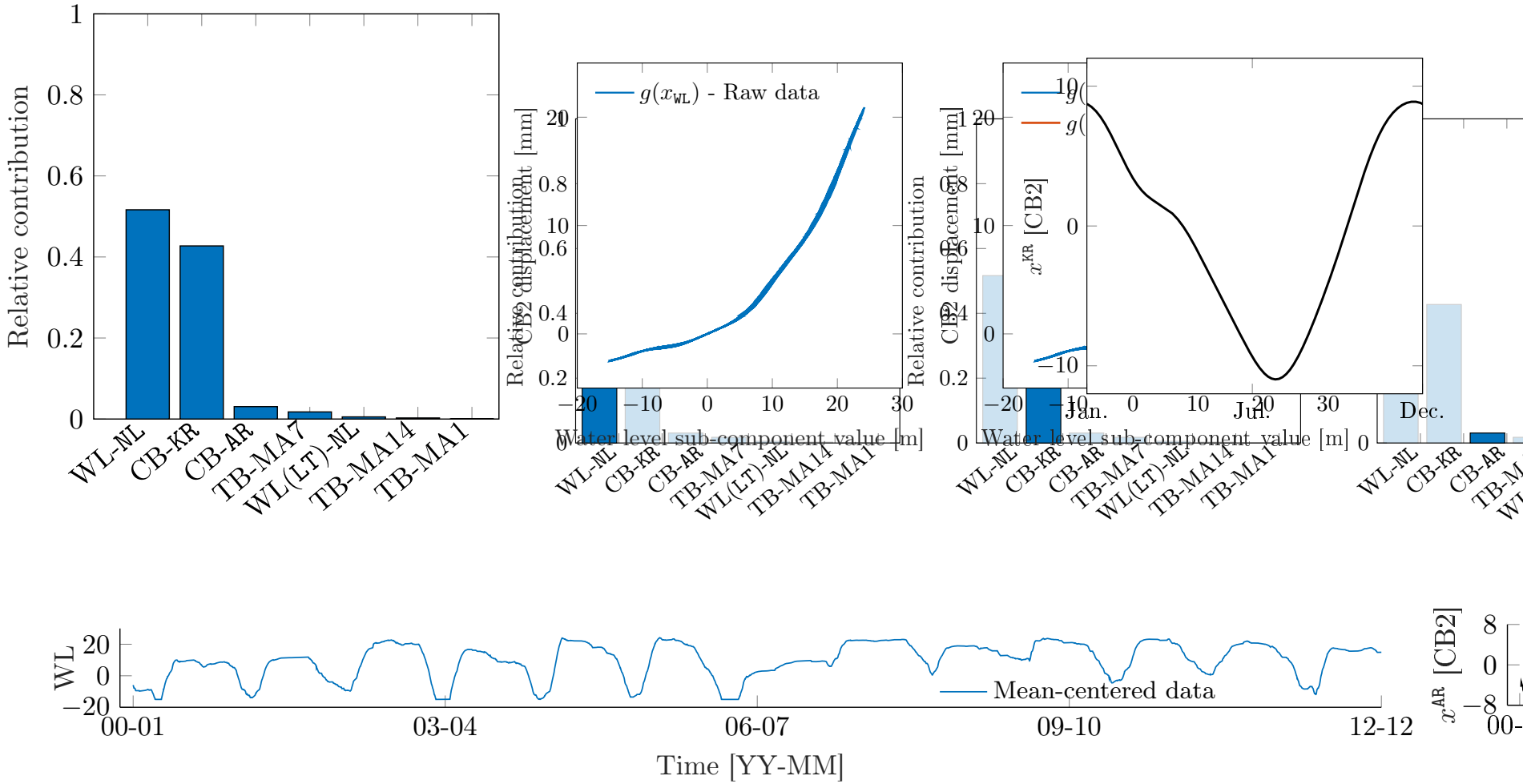
BDLM
($LL = -12.1$)



LSTM
($LL = -30.2$)



CB2 – BDLM's model structure & interpretation



LSTM v.s. BDLM...

BDLM

- Good accuracy ✓
- Uncertainties ✓
- Non-stationary data ✓
- Interpretable ✓
- Manual setup ✗

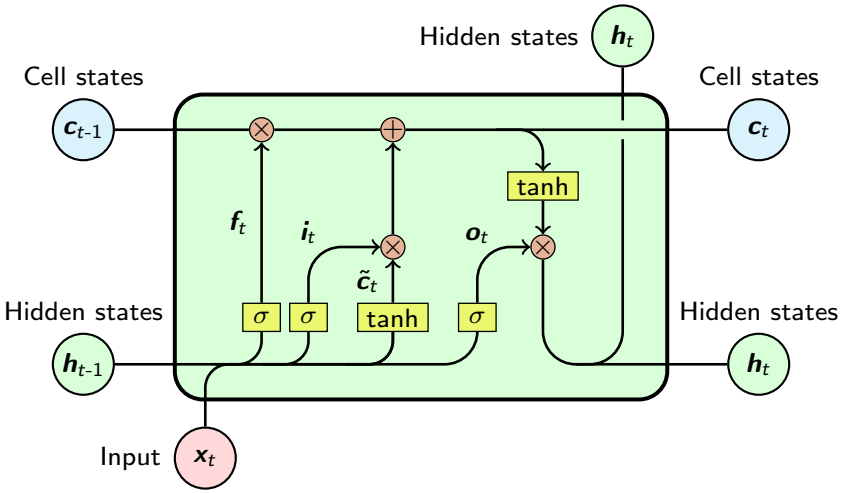
LSTM

- Good accuracy ✓
- Uncertainties ✓*
- Stationary data ✗
- Poor interpretability ✗
- Black-box setup ✓

Deka, Vuong, Goulet, Côté and Miquel, (2022). *Dam Behaviour Prediction Using an Ensemble of Bayesian Dynamic Linear Model and Bayesian LSTM Networks*, ICOLD workshop

Summary

$$\boldsymbol{c}_t = [c_1, c_2, \cdots, c_C]_t^T \in \mathbb{R}^C$$
$$\boldsymbol{h}_t = [h_1, h_2, \cdots, h_C]_t^T \in (-1, 1)^C$$



$\boldsymbol{f}_t$: Forget gate $\in (0, 1)^C$

$\boldsymbol{I}_t$: Input gate $\in (0, 1)^C$

$\tilde{\boldsymbol{c}}_t$: Candidate values $\in (-1, 1)^C$

$$\boldsymbol{c}_t = \underbrace{\boldsymbol{f}_t \odot \boldsymbol{c}_{t-1}}_{\text{forget}} + \underbrace{\boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t}_{\text{write}}$$

$\boldsymbol{o}_t$: Output gate $\in (0, 1)^C$

$$\boldsymbol{h}_t = \underbrace{\boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)}_{\text{Attention on a subset of the cell state}}$$