Leveraging machine learning emulators for efficient parameter estimation for 124 flux towers using SUMMA

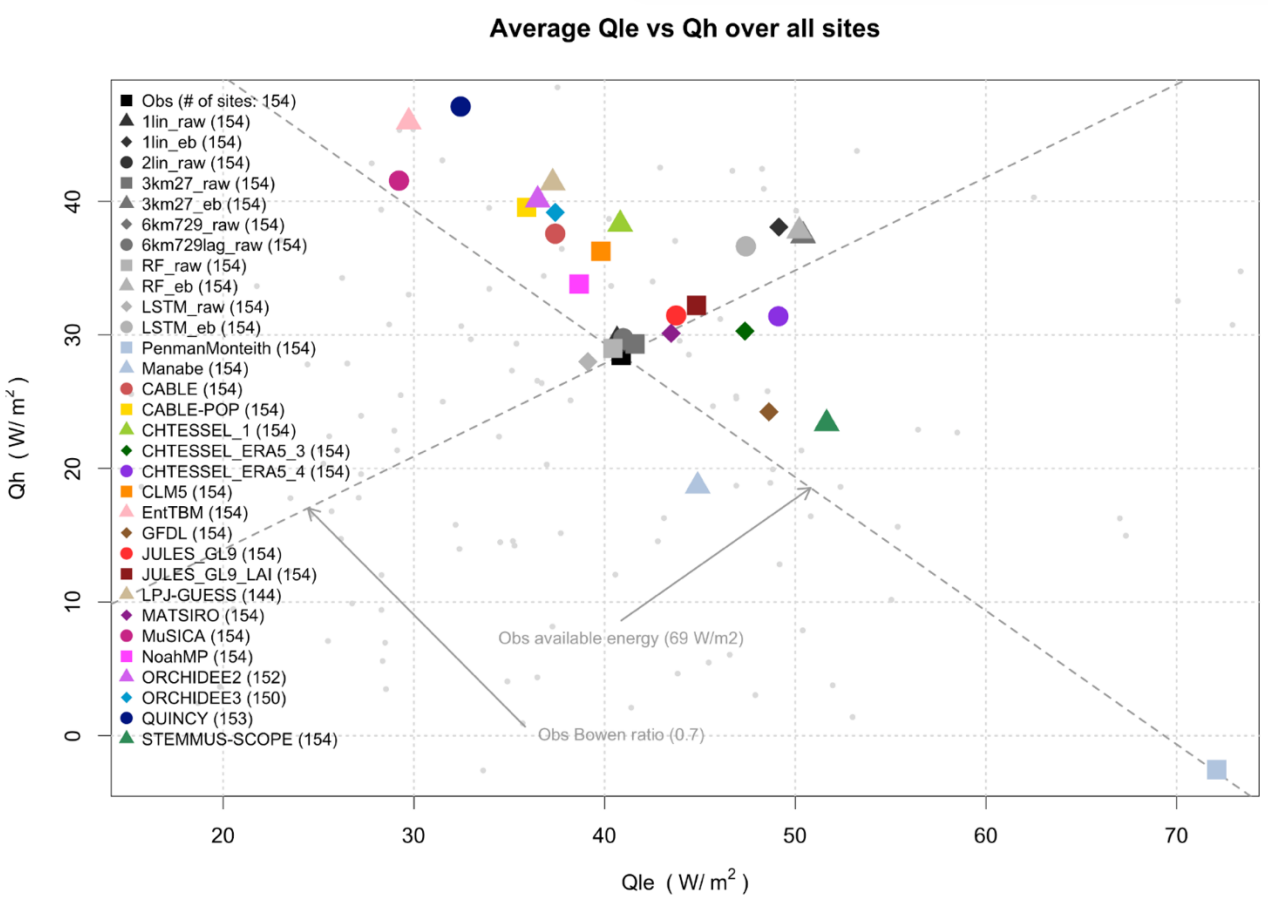
Ignacio Aguirre, Wouter Knoben, Nicolás Vásquez, and Martyn Clark
Department of Civil Engineering
University of Calgary, Alberta, Canada

NCAR | COMMUNITY EARTH SYSTEM MODEL CESM

2025 Workshop

Context

- Abramowitz et al. (2024) led the Plumber 2 MIP to evaluate the performance of models on turbulent fluxes (latent heat and sensible heat).
- It included 20 models and 7 benchmarks.
- Of the benchmarks,
  - Simplest: linear regression of the fluxes using short-wave radiation.
  - Complex: LSTM
- The models used default parameters (i.e., no calibration) and were evaluated using the entire period.

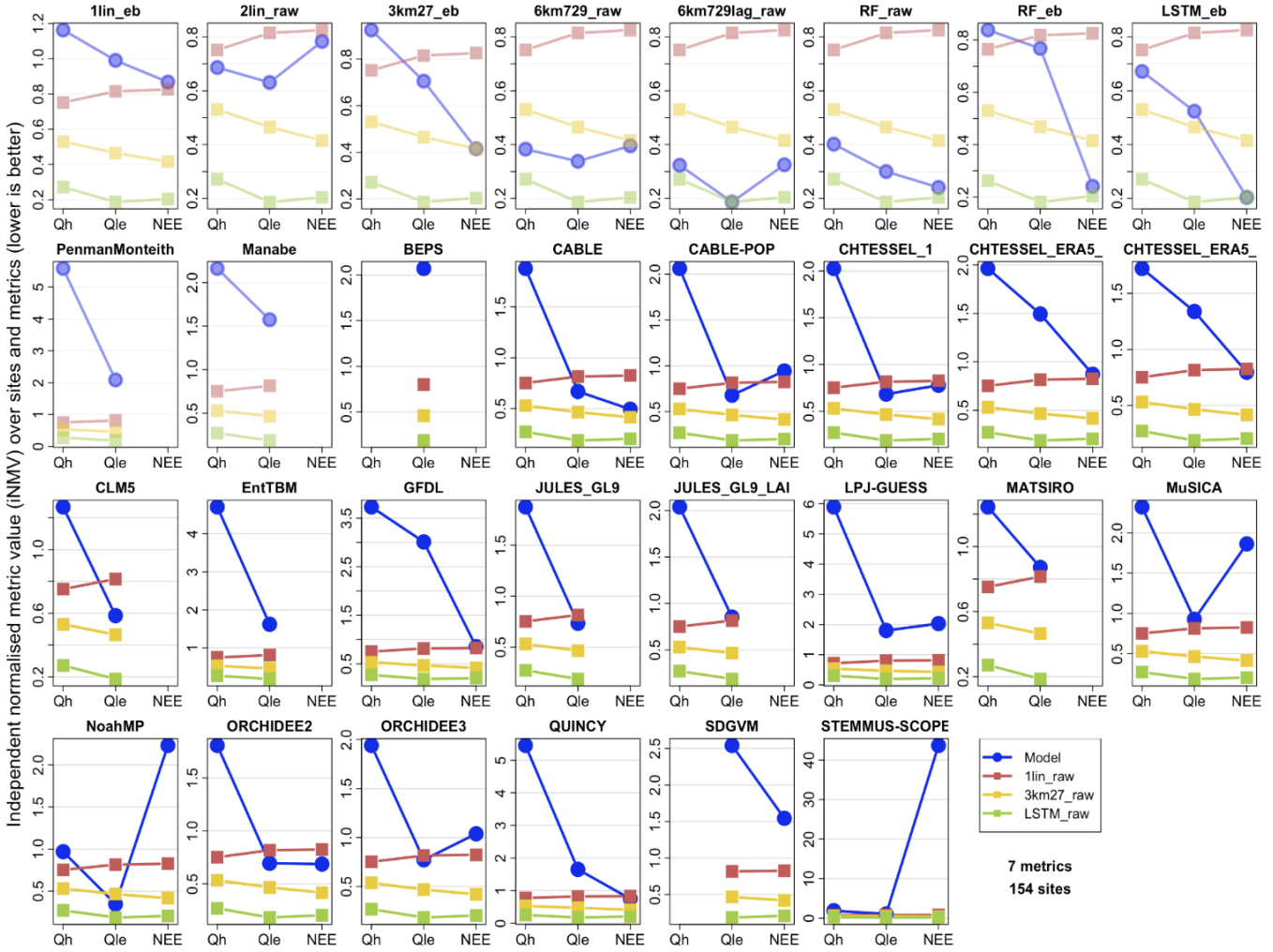


Abramowitz et al. (2024)

Context

Plumber 2 results show that:

- All state-of-the-art models are outperformed by simple regression for sensible heat (Qh)
- The LSTM benchmark for latent heat (Qle) outperforms all models
- There is no apparent relation between the level of complexity of the model and its performance.



Abramowitz et al. (2024)

Context

Abramowitz et al. (2024) concluded:

> This raises the question of whether LMs are too complex for the level of fidelity they provide. It's at least theoretically possible, for example, that an LM is perfect, but because we are unable to precisely prescribe its parameters for these site simulations (and global simulations) we are actively hindering its ability to get the right result. What the out-of-sample empirical models

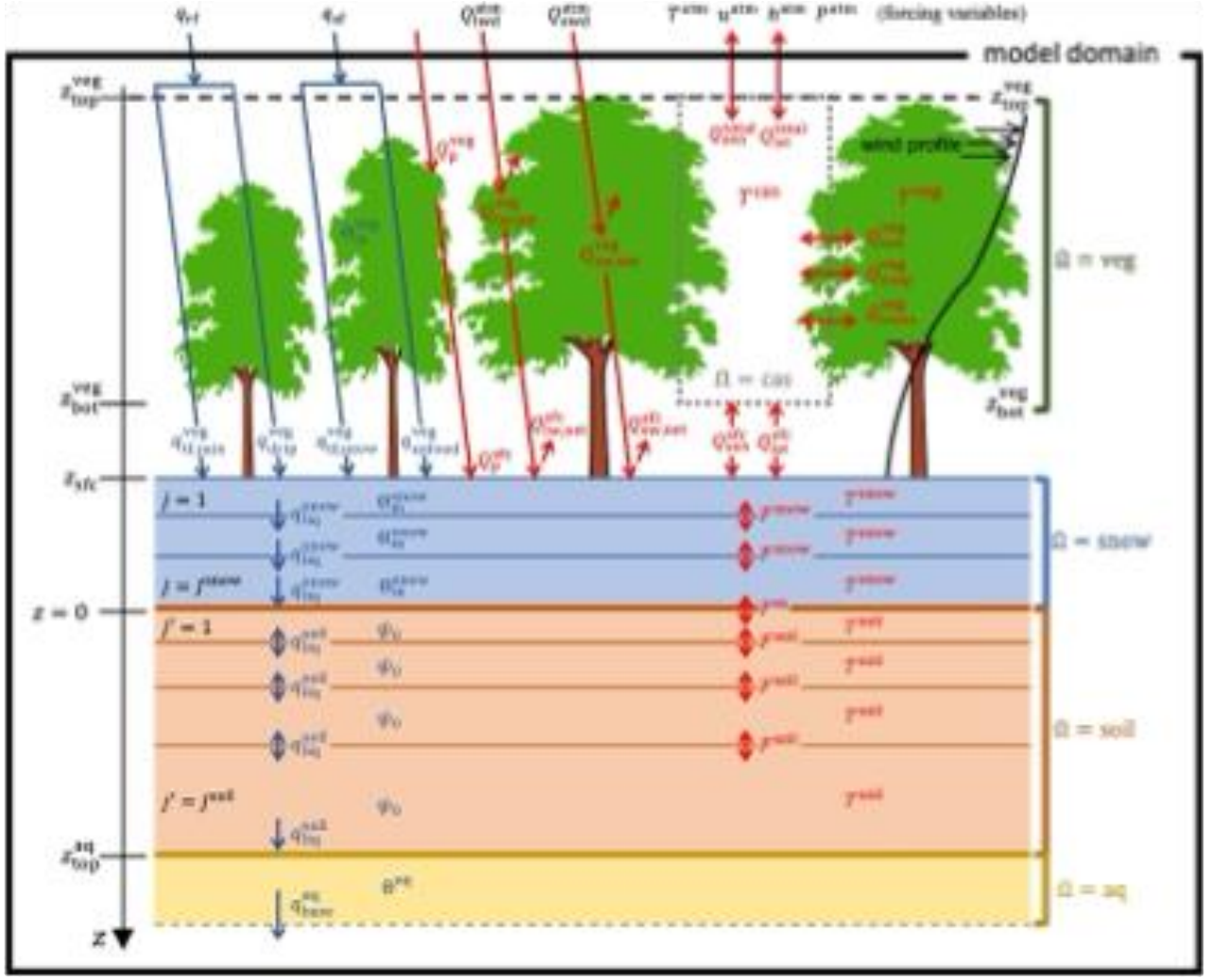
This study addresses this challenge and focuses on:

(a) Identify the key parameters that control Qle and Qh

(b) Estimate optimal parameters to improve the performance of Qh and Qle

Methodology

- To explore different modeling decisions, we utilized the Structure for Unifying Multiple Modeling Alternatives (SUMMA, Clark et al., 2015)
- **Simulate the conservation of mass and energy.**
- **Multiple modeling options for specific processes**
- Multiple state-of-the-art numerical solvers for the equations, including the SUNDIALS suite.
- Flexibility to adjust model parameters.
- Multiple options to represent horizontal and vertical heterogeneity.



Clark et al. (2021)

Methodology

- Sensitivity analysis (SA) and calibration were evaluated against measured-only fluxes (i.e., no-gap-filled or energy corrected).
- Data divided into:
  - Calibration: First 50% of the data
  - Evaluation: Second 50% of the data
- Objective metric: KGE

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

where:

r: is the Pearson correlation coefficient

α : is a term representing the variability of prediction errors

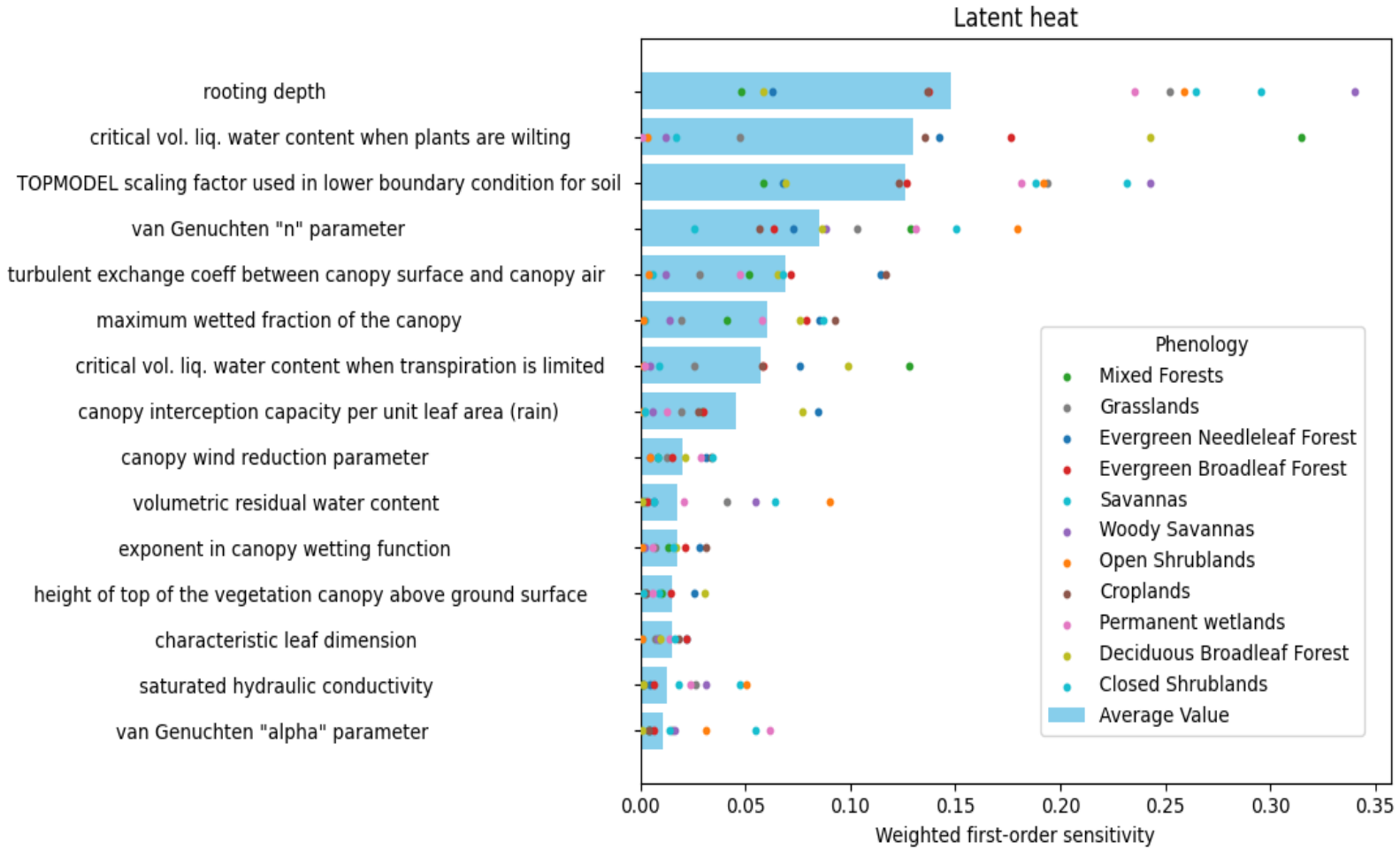
β: is a bias term

The plumber 2 towers were filtered from 170 to 124:

- Removed stations with less than 2 years
- Removed stations with unreliable observed data

Sensitivity analysis

1. Determined min and max bounds for each parameter.
2. Generated 4000 LHS samples to evaluate all 130 parameters.
3. Ran SUMMA.
4. Evaluated Qle and Qh
5. Determined first-order sensitivity using the variance-based method of PyVISCOUS (Liu et al., 2024)
6. Identified parameters that account collectively >85% of variance



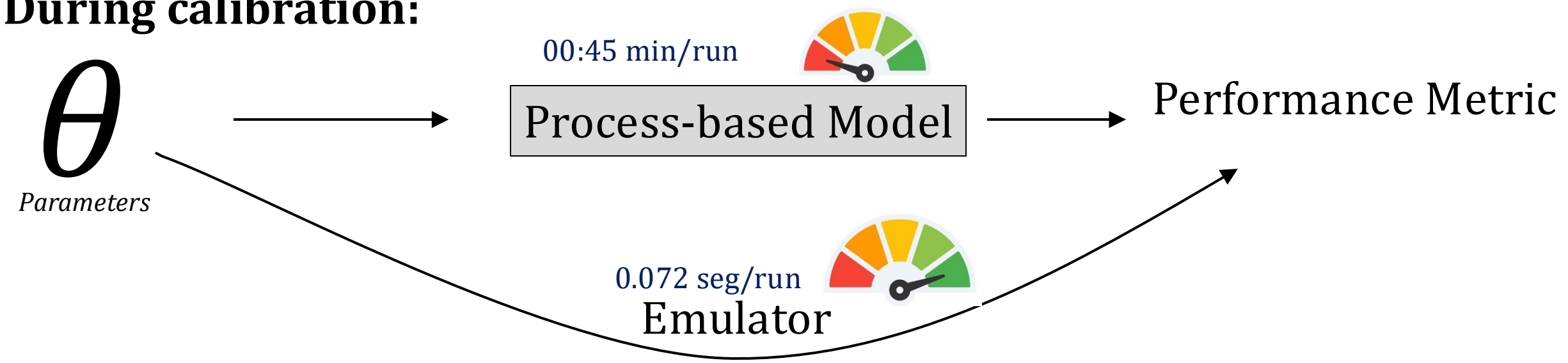
Calibration methodology: emulators

There are two kinds of emulators:

- Aim to reproduce the time-series of fluxes directly (i.e., by-passing the model parameters): Bennett et al. (2024); Maxwell et al. (2021).
- Aim to reproduce the dynamics of parameters and objective functions (i.e., not reproducing the timeseries of fluxes): Tang et al. (2024), Gong et al. (2016), and Herrera et al. (2022). **In this case, the model reproduces the fluxes, and the emulator is used as a surrogate during calibration.**

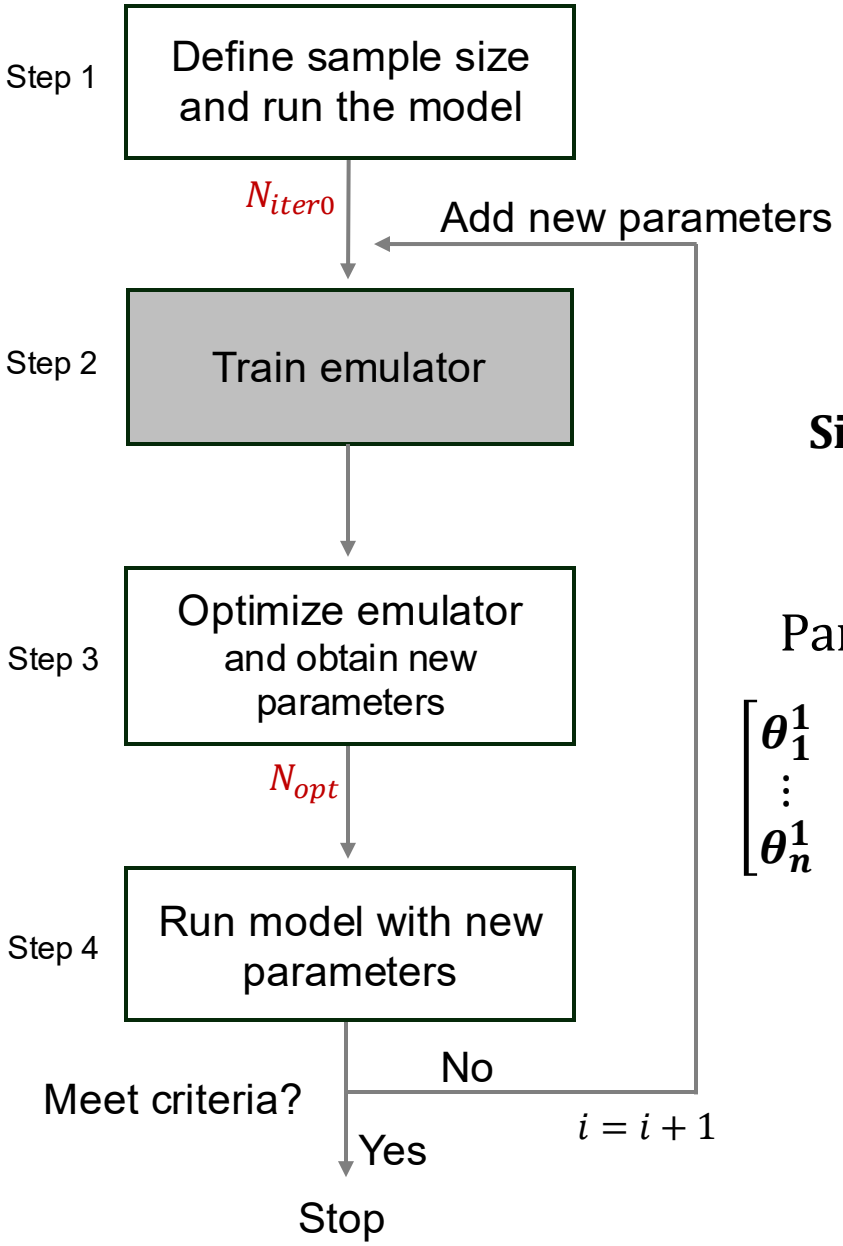
Final model evaluation is performed using the process-based model, not the emulator.

During calibration:



Calibration methodology

Based on the work of Guoqiang Tang, Andrew Wood, and Sean Swenson (2024, submitted to WRR)



Single site emulator

$$\boldsymbol{\theta}: \text{Parameters} \qquad \boldsymbol{y}: \text{Metric}$$

$$\begin{bmatrix} \boldsymbol{\theta_1^1} & \cdots & \boldsymbol{\theta_1^P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\theta_n^1} & \cdots & \boldsymbol{\theta_n^P} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{y_1} \\ \vdots \\ \boldsymbol{y_n} \end{bmatrix}$$

$$NKGE = \frac{KGE}{2 - KGE}$$

Large sample emulator

θ: Parameters $\quad A$: Attributes $\quad y$: Metric

$$\begin{bmatrix} \theta_1^1 & \cdots & \theta_1^P \\ \vdots & \ddots & \vdots \\ \theta_n^1 & \cdots & \theta_n^P \end{bmatrix} \begin{bmatrix} a_1^1 & \cdots & a_1^A \\ \vdots & \ddots & \vdots \\ a_1^1 & \cdots & a_1^A \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \text{ Tower 1}$$

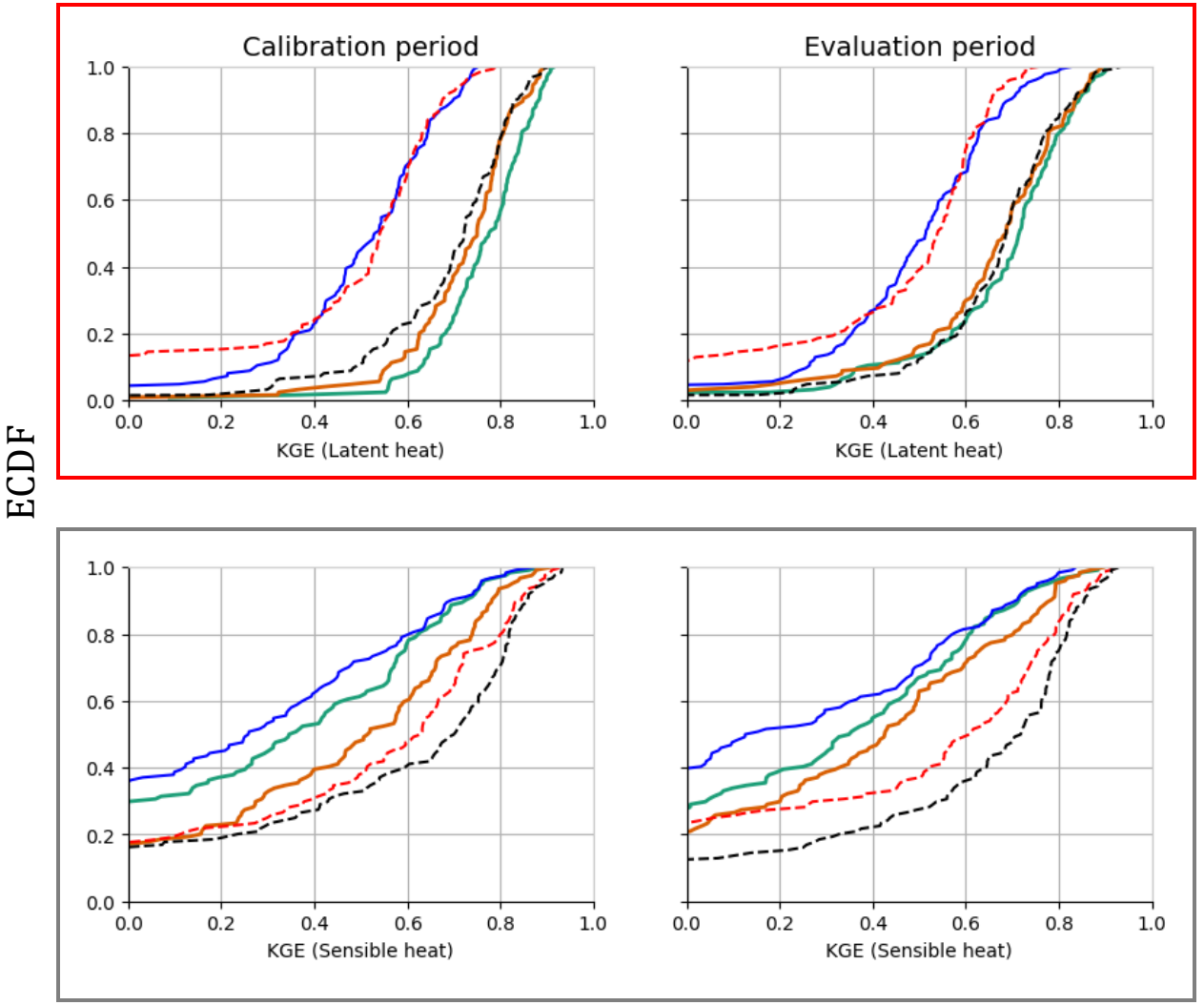
$$\begin{bmatrix} \theta_1^1 & \cdots & \theta_1^P \\ \vdots & \ddots & \vdots \\ \theta_n^1 & \cdots & \theta_n^P \end{bmatrix} \begin{bmatrix} a_2^1 & \cdots & a_2^A \\ \vdots & \ddots & \vdots \\ a_2^1 & \cdots & a_2^A \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \text{ Tower 2}$$

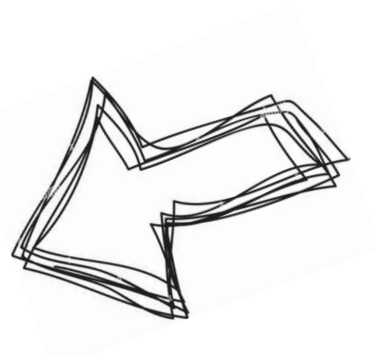
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$\begin{bmatrix} \theta_1^1 & \cdots & \theta_1^P \\ \vdots & \ddots & \vdots \\ \theta_n^1 & \cdots & \theta_n^P \end{bmatrix} \begin{bmatrix} a_B^1 & \cdots & a_B^A \\ \vdots & \ddots & \vdots \\ a_B^1 & \cdots & a_B^A \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \text{ Tower B}$$

SSE: Objective latent heat (Qle)

- SSE (Qle) outperforms the linear regression and LSTM benchmark for Qle.
- Selecting the best simulation is key and can be performed on:
  - Only the flux that is being calibrated
  - **Both fluxes.**





The focus of this experiment is latent heat.

SSE results

- SSE (Qle) based on highest KGE (Qle)
- SSE (Qle) based on highest KGE (Qle & Qh)

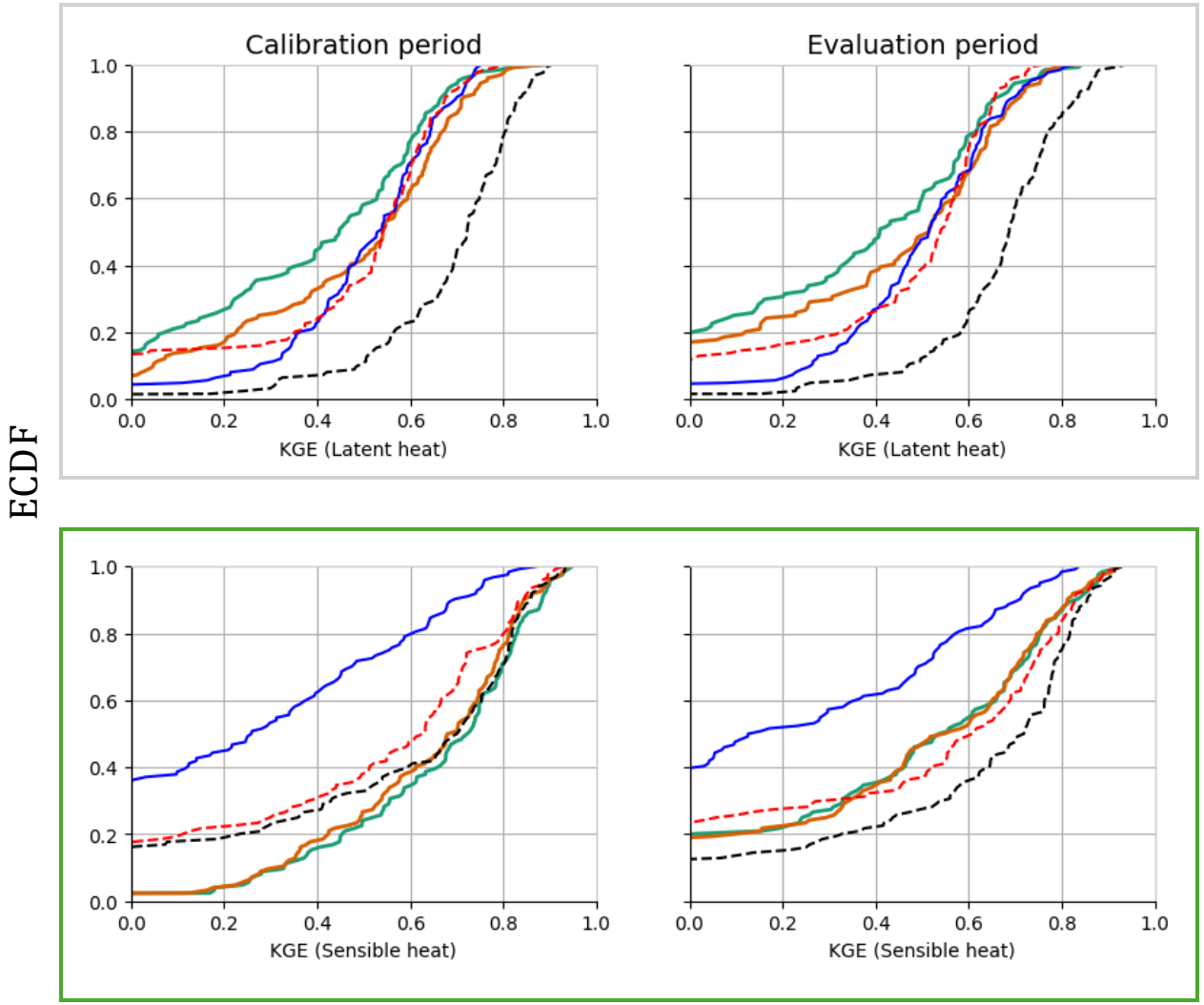
Reference

- SUMMA default
- Linear regression ref. (Plumber 2)
- LSTM ref. (Plumber 2)

These are the result of sensible heat, while calibrating latent heat

SSE: Objective sensible heat (Qh)

- SSE (Qh) outperforms the simple regression and LSTM reference in the calibration period, but not during the evaluation period.
- The performance of SSE (Qh) in latent heat is limited.



These are the result of latent heat, while calibrating sensible heat

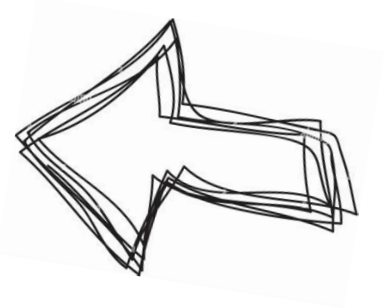
SSE results

- SSE (Qh) based on highest KGE (Qh)
- SSE (Qh) based on highest KGE (Qle & Qh)

Reference

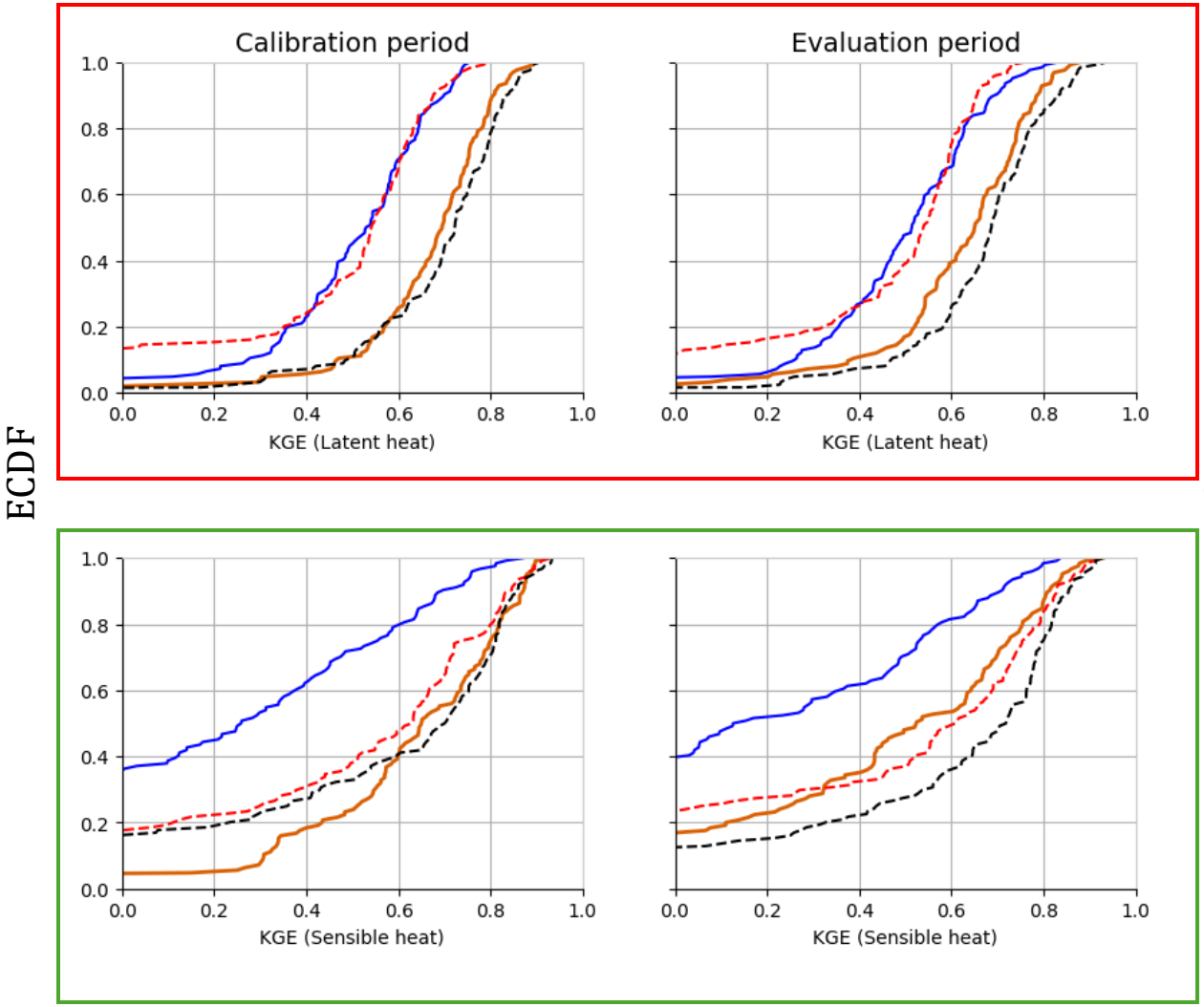
- SUMMA default
- Linear regression ref. (Plumber 2)
- LSTM ref. (Plumber 2)

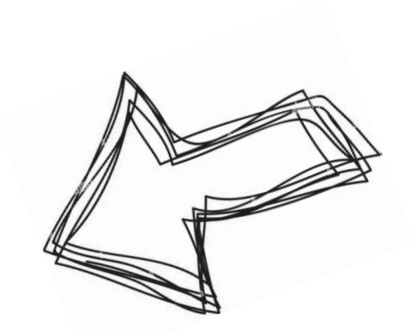
The focus of this experiment is sensible heat.



SSE: Objective latent and sensible heat

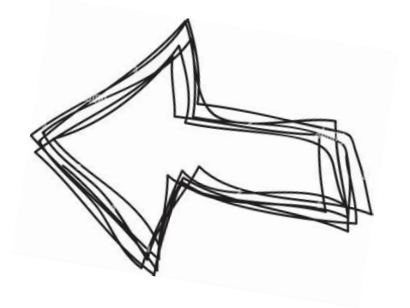
- SSE (Qle & Qh) was built weighing each flux 50%
- It yields performance comparable to LSTM for Qle and similar to linear regression for Qh.





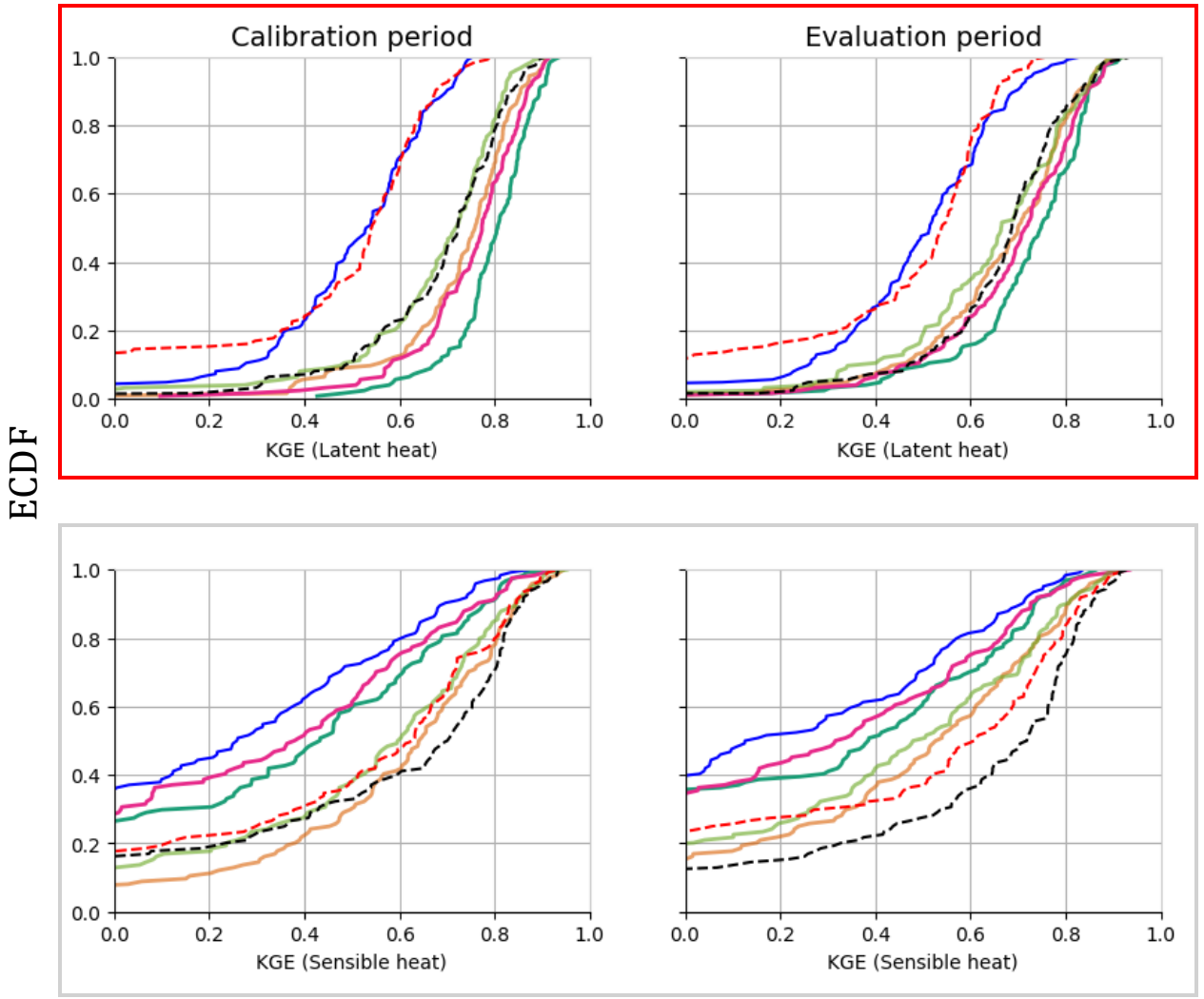
The focus of this experiment is on both latent heat and sensible heat

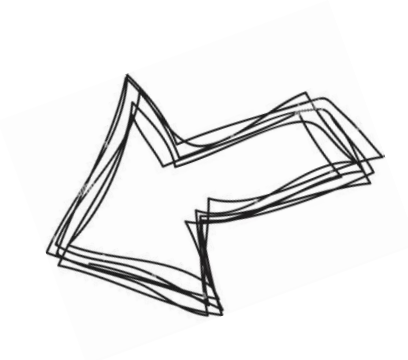
The focus of this experiment is on both latent heat and sensible heat



LSE: Objective latent heat

- LSE (Qle) outperforms the linear regression and LSTM references.
- We evaluated using ~50 parameters and ~30 parameters:
  - Simulations with ~50 parameters have better performance in KGE (Qle) only and KGE (Qle & Qh)
  - The emulator approach does not exhibit saturation with 50 parameters. The point of saturation (i.e., achieving the same performance with more parameters) has not been defined yet.





The focus of this experiment is latent heat.

LSE results

- LSE (Qle) | highest KGE (Qle) | ~ 50 params
- LSE (Qle) | highest KGE (Qle & Qh | ~ 50 params
- LSE (Qle) | highest KGE (Qle) | ~ 30 params
- LSE (Qle) | highest KGE (Qle & Qh | ~ 30 params

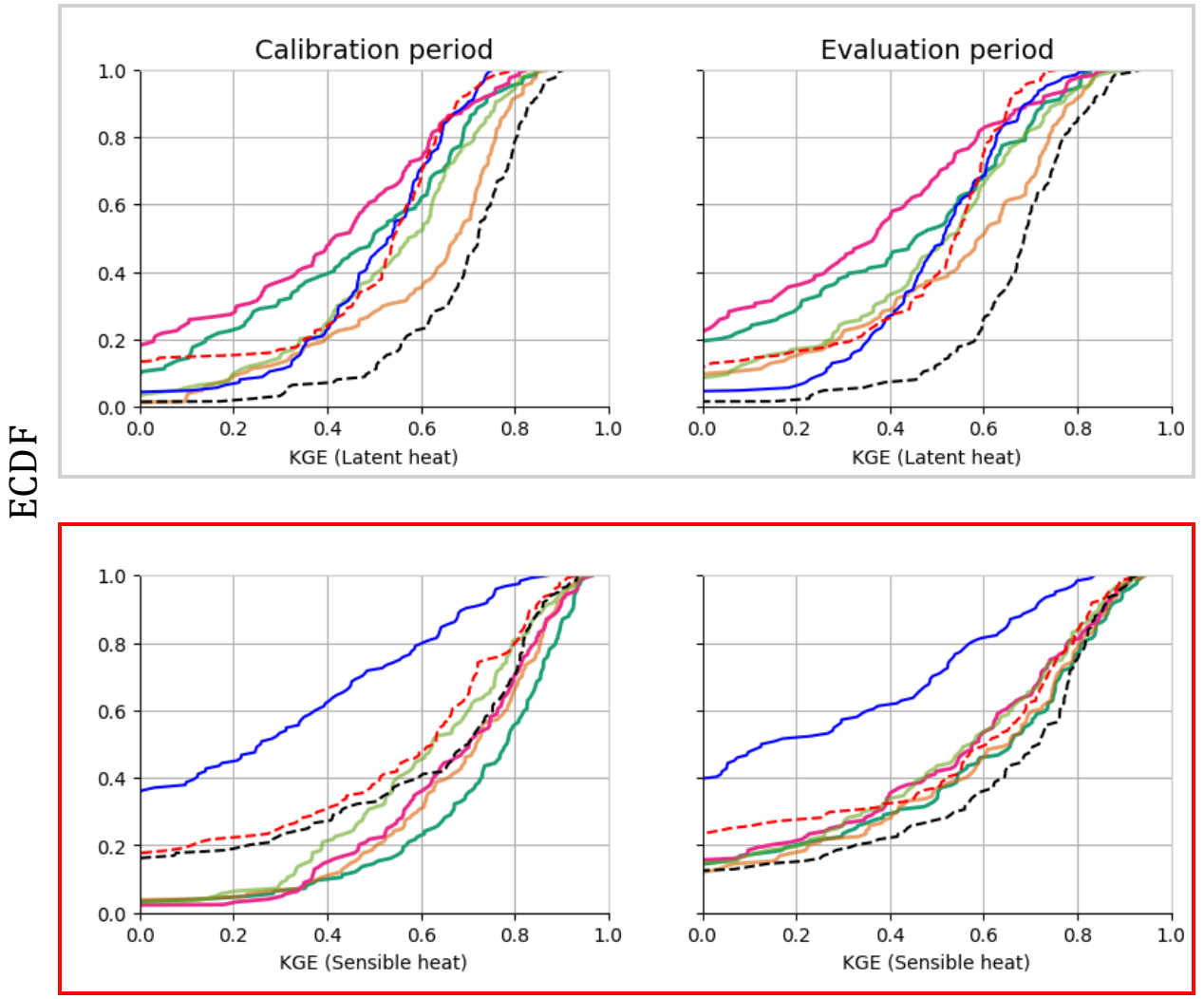
Reference

- SUMMA default
- Linear regression ref. (Plumber 2)
- LSTM ref. (Plumber 2)

These are the result of sensible heat, while calibrating latent heat

LSE: Objective sensible heat

- LSE (Qh) outperforms the references on the calibration period, but not in the evaluation period.
- The results align with Plumber 2, which shows that models are limited in capturing Qh.
- As observed in LSE (Qle), the LSE with ~50 parameters outperforms the LSE with ~30 parameters.



These are the result of latent heat, while calibrating sensible heat

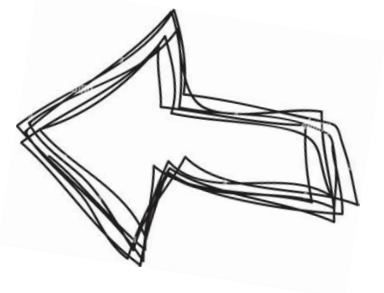
LSE results

- LSE (Qh) | highest KGE (Qh) | ~ 50 params
- LSE (Qh) | highest KGE (Qle & Qh | ~ 50 params
- LSE (Qh) | highest KGE (Qh) | ~ 30 params
- LSE (Qh) | highest KGE (Qle & Qh | ~ 30 params

Reference

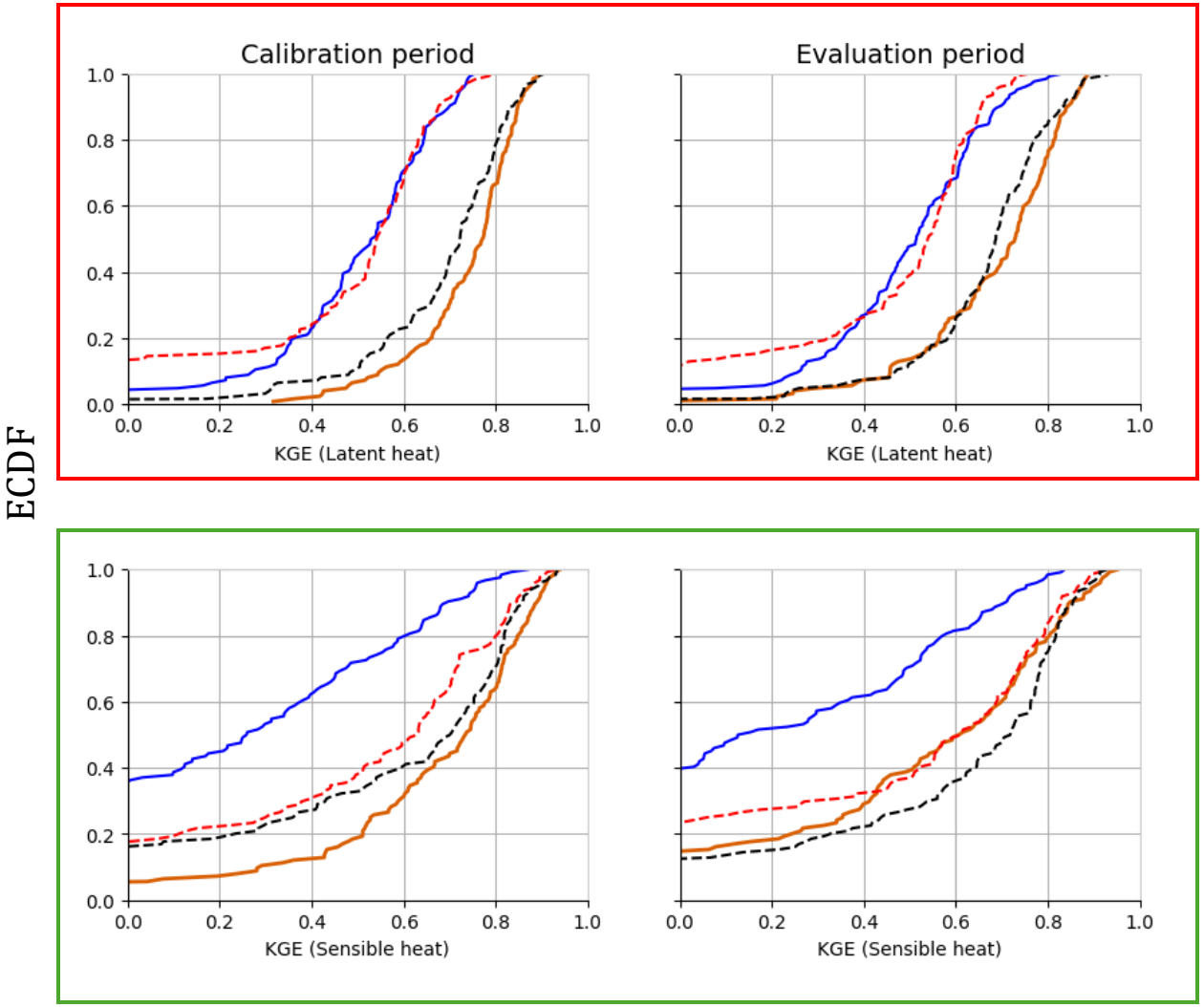
- SUMMA default
- Linear regression ref. (Plumber 2)
- LSTM ref. (Plumber 2)

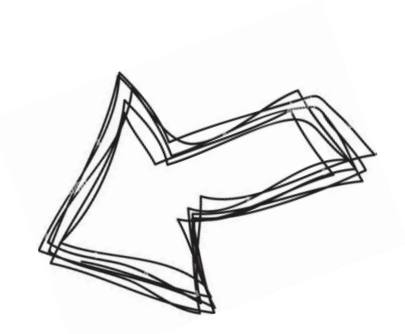
The focus of this experiment is sensible heat.



LSE: Objective latent and sensible heat

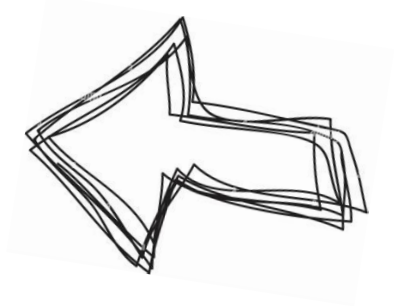
- LSE (Qle & Qh) was built weighing each flux 50%
- For latent heat, it outperforms the LSTM on the calibration and validation periods.
- For the sensible heat, it outperforms the benchmarks only in calibration, similar to LSE(Qh).





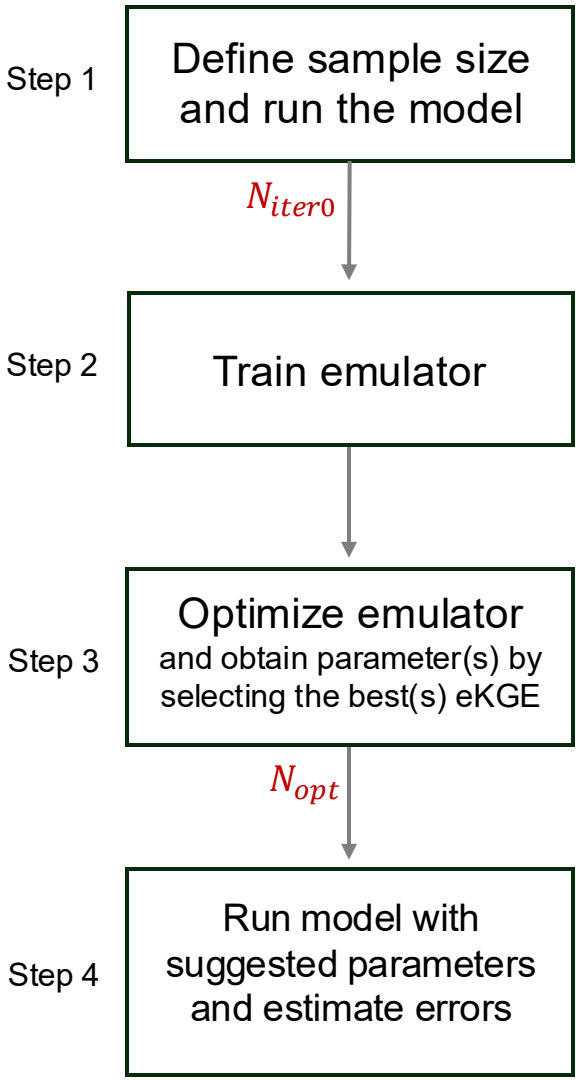
The focus of this experiment is on both latent heat and sensible heat

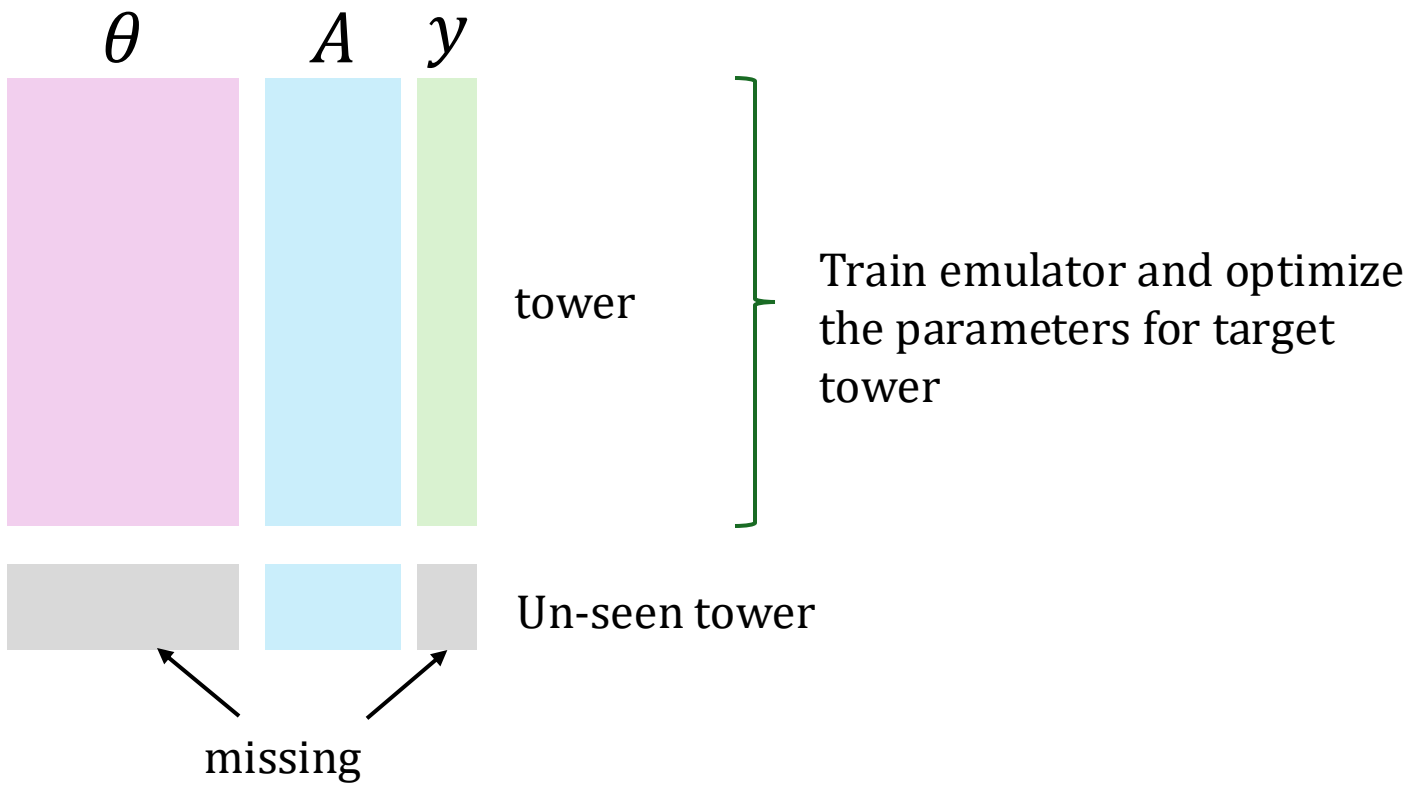
The focus of this experiment is on both latent heat and sensible heat



Calibration methodology: regionalization

Test the emulator on flux towers not included in the training data





Based on the work of Guoqiang Tang, Andrew Wood, and Sean Swenson (2024, submitted to WRR)

LSE applied to regionalization (Obj: Qle)

To select the best performance:

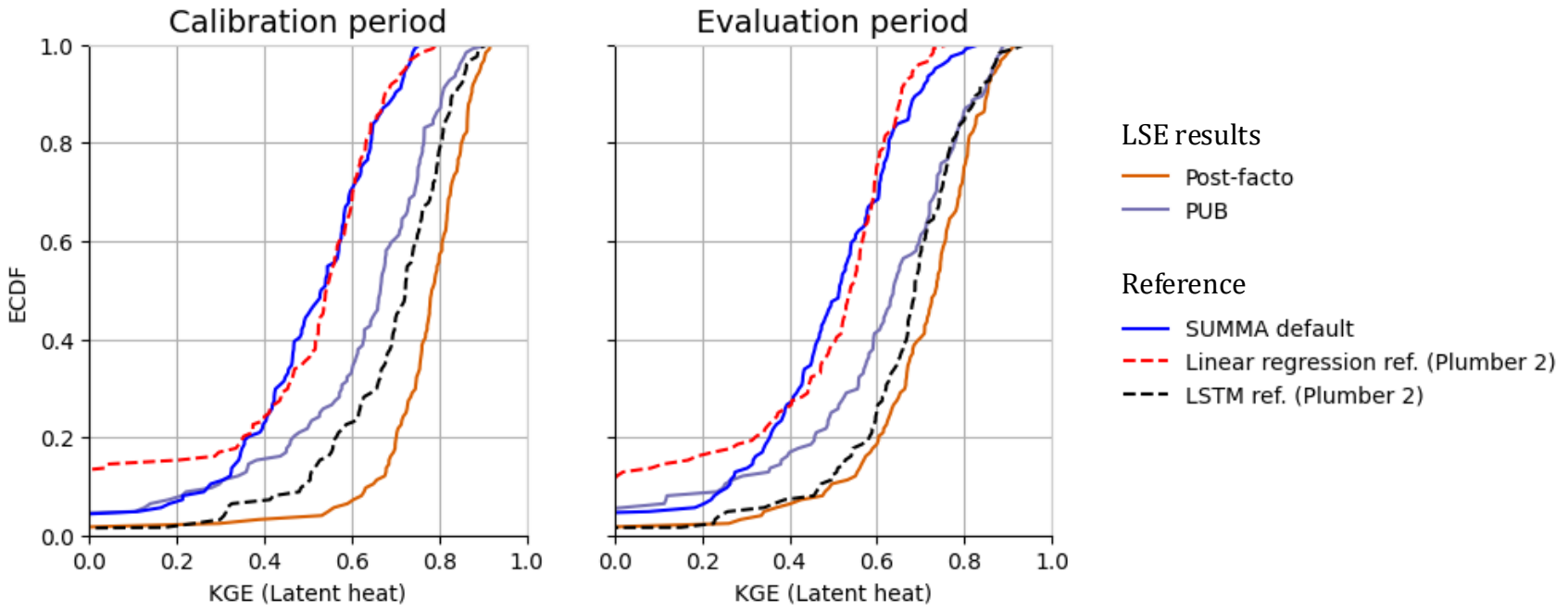
- Check the best emulated metric: PUB approach.
- Check the best metric from the hydrological model post-facto analysis

These two options differ in the error between the KGE estimated by the emulator vs the KGE estimated by SUMMA.

The emulator often estimates higher values than SUMMA, which has also been observed in Tang et al. (2024, submitted to WRR) and Farahani et al. (2024, submitted to HESS)

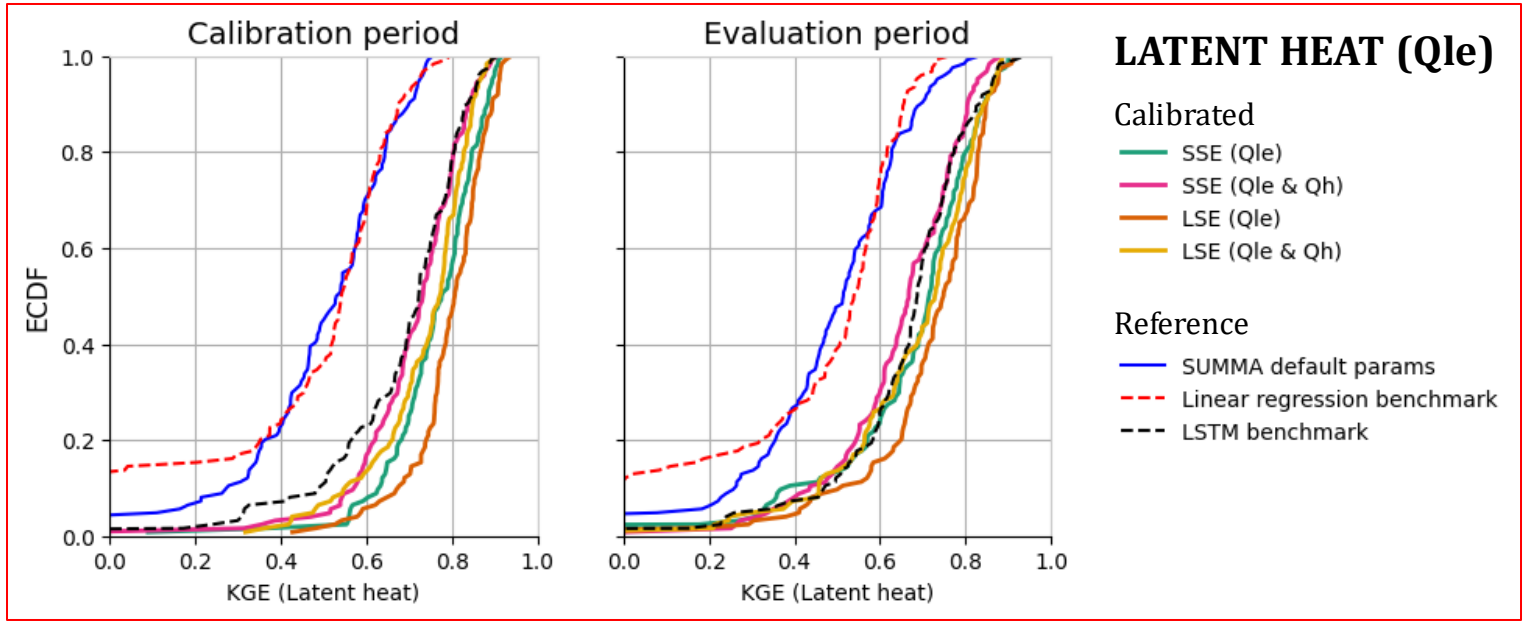
G1: 20% of towers	G2: 20% of towers	G3: 20% of towers	G4: 20% of towers	G5: 20% of towers

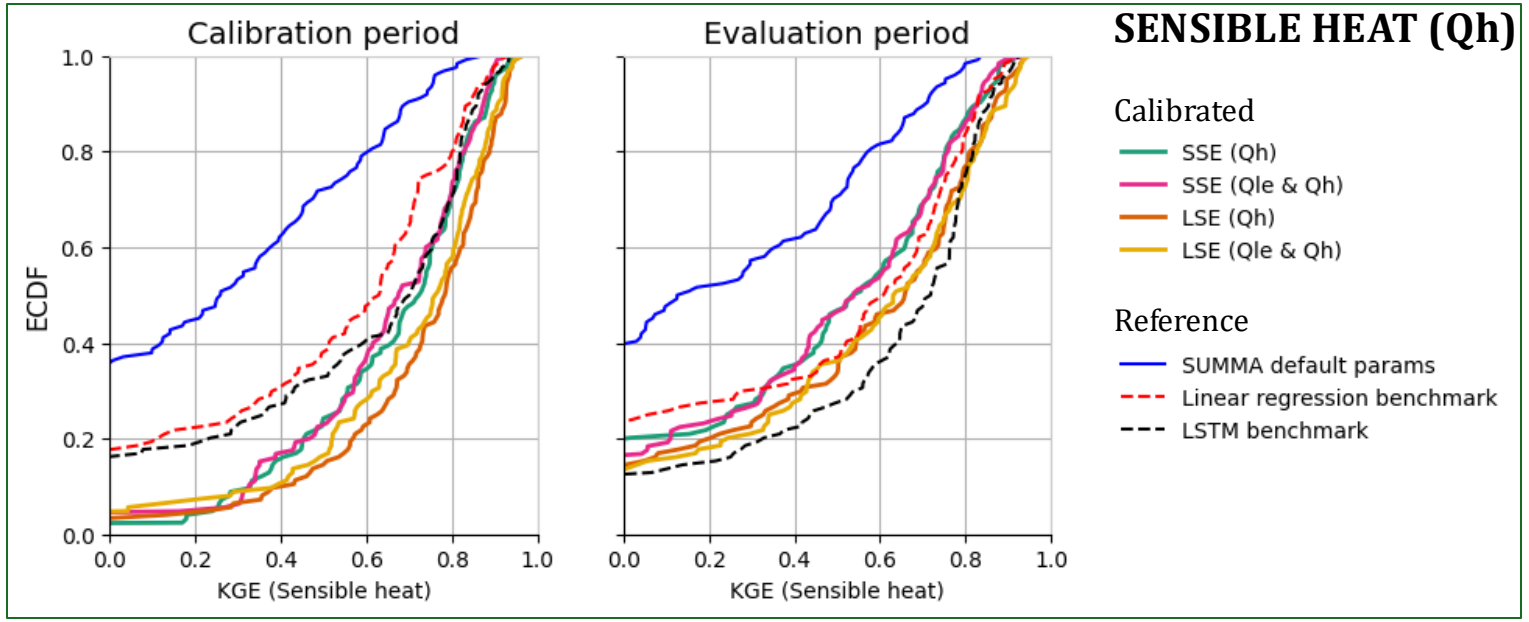
Train the emulator on 80% of the towers, test on the 20% unseen by the emulator



Summary

- Qle and Qh calibrated can outperform reference benchmarks.
- LSE outperforms SSE for Qle and Qh.
- LSE trained with ~50 parameters outperforms LSE (~30 params)
- The LSE method provides good results for regionalization. Thus, the emulator can be used to find parameters for towers not included in the training dataset.
- There are limitations on capturing Qh.

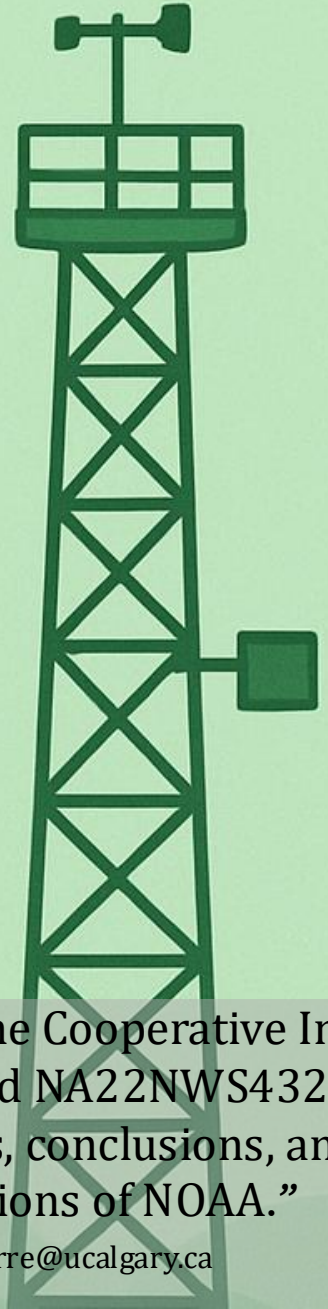




Future work

- Using SUMMA:
  - Evaluate the impact of the multiple decisions used in the emulator (e.g., number of iterations)
  - Evaluate more applications of regionalization, such as train in 20%, and evaluate over 80%
  - Evaluate ways to improve results on sensible heat
- Apply this method to the same PLUMBER2 towers using CLM5 (Lawrence et al., 2019)

Questions



Acknowledgments

- Kyle Klenk, Ashley Van Beusekom, Darri Eythorsson, Befekadu Woldegiorgis, Kasra Keshavarz, Peter Wagener, Dave Lawrence, Andrew Wood, and Gordon Bonan.

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the opinions of NOAA."