Benchmarking machine learning-based emulators and traditional methods to calibrate land model parameters for 124 global flux tower sites

Ignacio Aguirre[1], Wouter Knoben[1], Nicolás Vásquez[1], & Martyn Clark[1]

[1] Schulich School of Engineering. University of Calgary, Canada

EGU26-14172

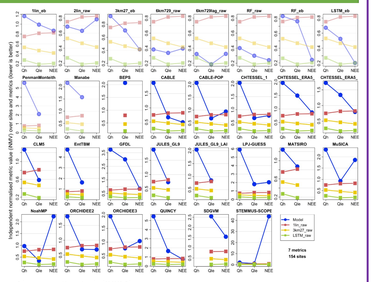
This presentation participates in OSPP

Outstanding Student & PhD candidate Presentation contest

Background

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.
Plumber 2 results show that:

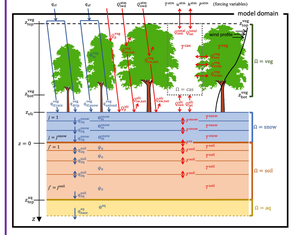
- All state-of-the-art models are outperformed by simple regression for sensible heat (Qh)
- The LSTM benchmark for latent heat (Qle) outperforms all models
- As in the catchment hydrology example by Nearing et al. (2021), they show that LSTM performs better than process-based models; however, model performance can be improved through calibration.



Abramowitz et al. (2024)

This study addresses this challenge through calibration, comparing different calibration methods (e.g., single-site emulator, large-sample emulator, genetic algorithm, and dimensional search) to identify an optimal parameter set for latent heat and sensible heat, and evaluating its performance using temporal and spatial cross-validation.

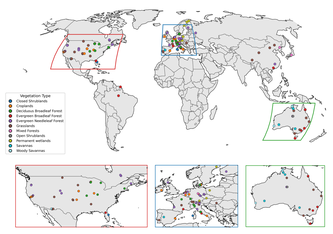
Methodology



To represent hydrological processes, we utilized the process-based model of 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.

To capture different hydrological regimes, we used the PLUMBER 2 dataset (Abramowitz et al., 2023; Ukkola et al., 2022).
We removed stations with data issues or fewer than 2 years of data, resulting in 124 flux towers.
We divided the data into calibration (1st 50% of the data) and validation (2nd 50% of the data)
We only used the measured-only data for latent heat and sensible heat to compute the metrics.



Using the Kling-Gupta efficiency (KGE, Kling et al., 2012) and focusing on a composed metric of 50% of latent heat (Qle) and 50% of sensible heat (Qh). We compared the default SUMMA runs against four calibration methods and two data-driven benchmarks.
These are the calibration methods:

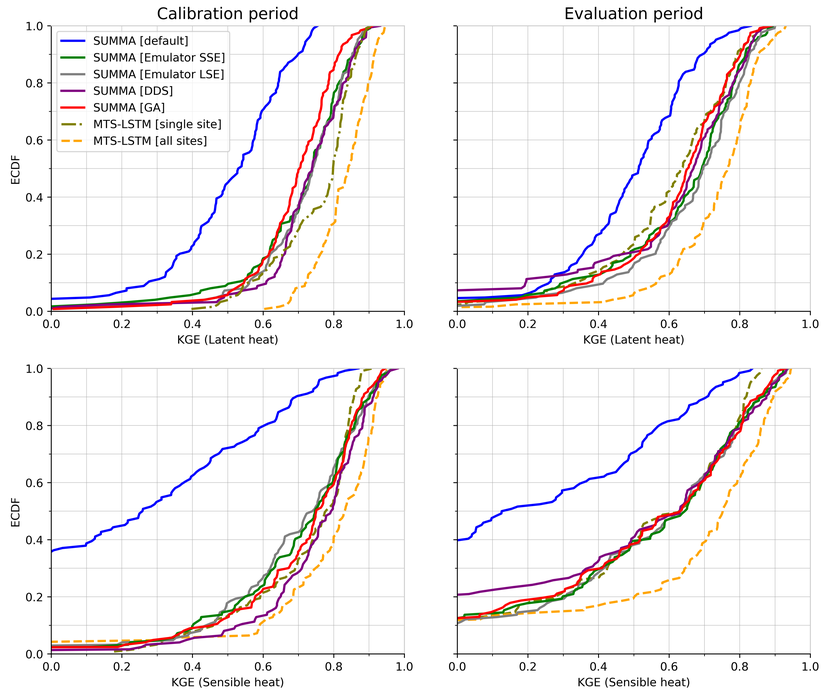
1. Calibration using the single-site emulator (SSE; Tang et al., 2025), where a machine-learning emulator is trained only with data from one site. The initial data consist of 600 runs, and the emulator is run for 17 iterations, resulting in 2300 simulations.
2. Calibration using the large-sample emulator (LSE; Tang et al., 2025), where a machine-learning emulator is trained using the data from the 124 flux towers at the same time. The initial data consist of 600 runs, and the emulator is run for 20 iterations, resulting in 2600 simulations.
3. Calibration using the Dynamically Dimensioned Search algorithm (DDS; Tolson and Shoemaker, 2007) using the implementation in the Ostrich software. We used 2000 simulations.
4. Calibration using a genetic algorithm (GA; Yoon and Shoemaker, 2001) using the implementation in Ostrich. We also used 2000 simulations.

Plus, we used two benchmarks based on Gauch et al. (2021) and Kratzert et al. (2022).

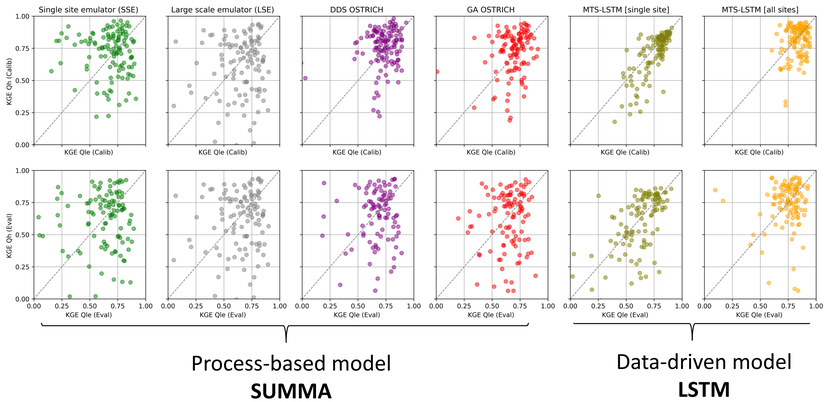
1. Lower-bound benchmark: a multifrequency (daily and 30-minute) MTS-LSTM (Gauch et al., 2021; Kratzert et al., 2022) trained using data from a single site. Key hyperparameters include an output dropout of 0.4, 100 training epochs, and a hidden size of 64.
2. Upper-bound benchmark: a multifrequency (daily and 30-minute) MTS-LSTM (Gauch et al., 2021; Kratzert et al., 2022) trained jointly on data from 124 sites. Key hyperparameters include an output dropout of 0.4, 30 training epochs, and a hidden size of 64

Results

The emulator methods (LSE and SSE) achieve performance comparable to traditional methods (DDS and GA) during the calibration period and outperform them during the validation period.

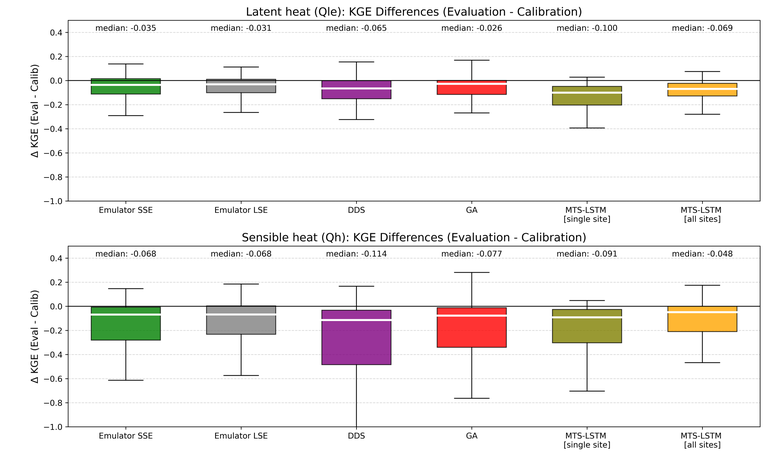


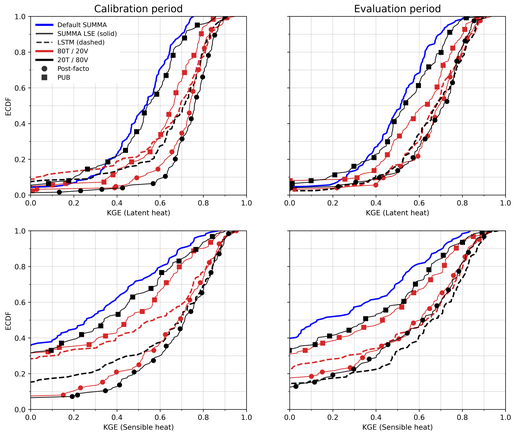
- Different calibration methods can significantly improve the performance of process-based models.
- **Calibration methods and benchmarks can, at the same time, improve the representation of latent heat (Qle) and sensible heat (Qh).**



Results

Looking at the performance in calibration and validation periods: The emulator LSE can achieve the lowest overall differences between the two periods, ensuring consistent temporal validation.





The emulator LSE can also be used to identify parameters in unseen flux towers, as demonstrated by the spatial cross-validation.

- These are results trained with 20% of the basins and tested on the remaining 80% and vice versa.
- The LSTM was built using multi-frequency MTS-LSTM (Gauch et al., 2021; Kratzert et al., 2022) using 30 epochs, a hidden size of 64, and an output dropout of 0.4.
- The LSE results can be evaluated using the simulated KGE (PUB approach) or using the actual KGE values (Post-facto).

Future work

- Use an emulator trained on flux tower data to identify parameter sets that best reproduce evapotranspiration patterns across North America and the planet (regionalization).
- Evaluate the impact of different model configurations (different modeling equations and parameter values), testing both default and optimized parameters to minimize errors in latent and sensible heat fluxes.
- Conduct hyperparameter tuning to define a diverse set of LSTM-based benchmarks.

Contact:
Ignacio Aguirre ignacio.aguirre@ucalgary.ca
Wouter Knoben wouter.knoben@ucalgary.ca
Nicolas Vasquez nicolas.vasquez@ucalgary.ca
Martyn Clark martyn.clark@ucalgary.ca

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.