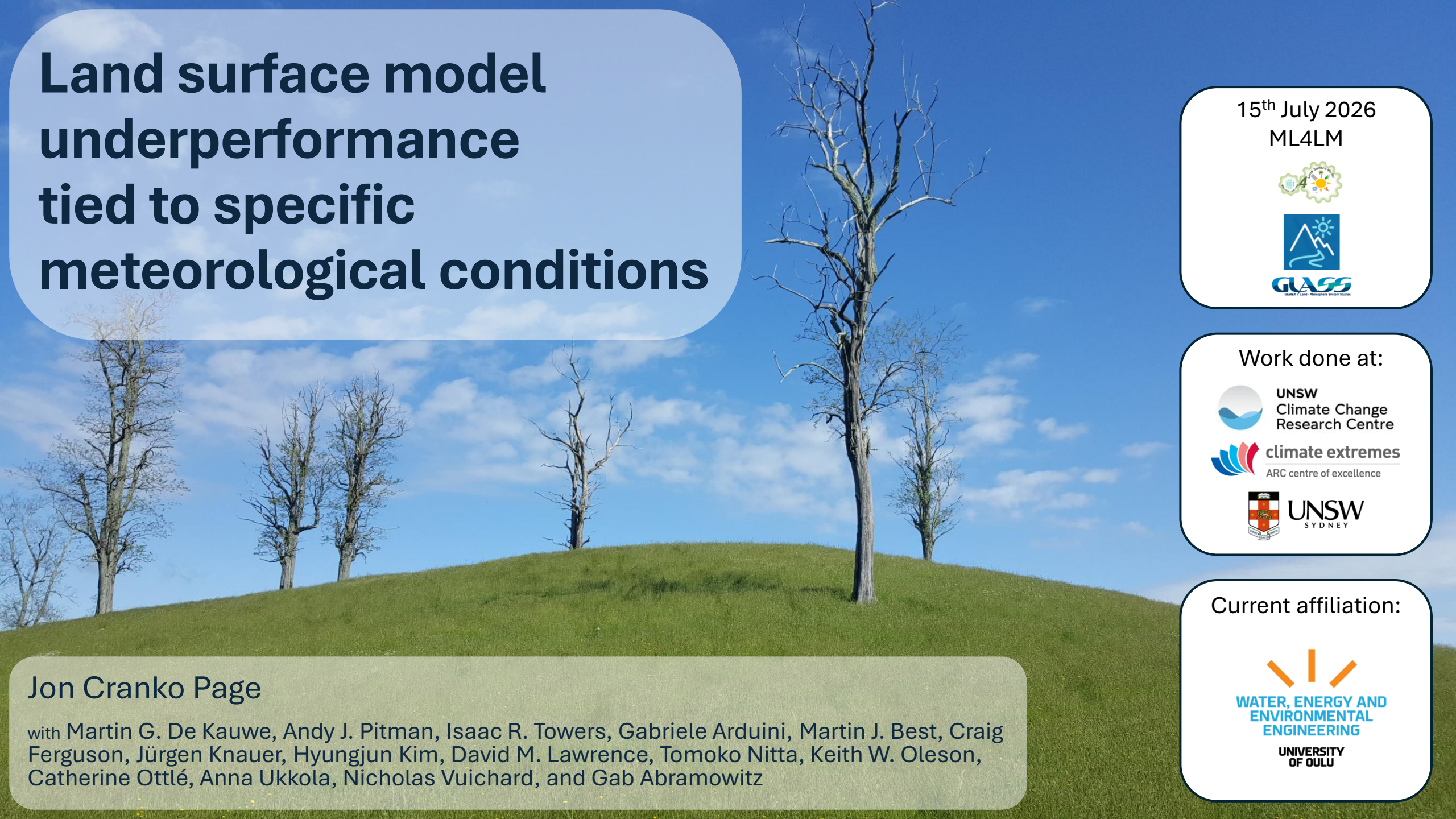
Land surface model underperformance tied to specific meteorological conditions



Jon Cranko Page

with Martin G. De Kauwe, Andy J. Pitman, Isaac R. Towers, Gabriele Arduini, Martin J. Best, Craig Ferguson, Jürgen Knauer, Hyungjun Kim, David M. Lawrence, Tomoko Nitta, Keith W. Oleson, Catherine Ottlé, Anna Ukkola, Nicholas Vuichard, and Gab Abramowitz

15th July 2026
ML4LM

Work done at:

UNSW Climate Change Research Centre

climate extremes
ARC centre of excellence

UNSW SYDNEY

Current affiliation:

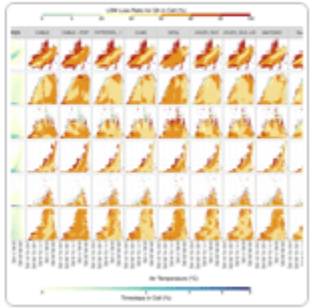
WATER, ENERGY AND ENVIRONMENTAL ENGINEERING
UNIVERSITY OF OULU

Our study based on the **benchmarking framework PLUMBER2**
identified that a collection of **land surface models underperformed**
simulating carbon, water, and energy fluxes **relative to empirical benchmarks**
when **two (or more) meteorological drivers** were **comparatively extreme**

Research article | 09 Jan 2026

Land surface model underperformance tied to specific meteorological conditions

Jon Cranko Page, Martin G. De Kauwe, Andy J. Pitman, Isaac R. Towers, Gabriele Arduini, Martin J. Best, Craig R. Ferguson, Jürgen Knauer, Hyungjun Kim, David M. Lawrence, Tomoko Nitta, Keith W. Oleson, Catherine Ottlé, Anna Ukkola, Nicholas Vuichard, Xiaoni Wang-Faivre, and Gab Abramowitz



When **meteorological conditions** are at the **edge of the joint distribution** of the input domain space,
our **land surface models should be doing better** with the information they are provided.

What is benchmarking?

*Meaning of **benchmarking** in English*

benchmarking

noun [U]

UK /ˈbentʃˌmɑː.kɪŋ/ US /ˈbentʃˌmɑːr.kɪŋ/

Add to word list

the act of measuring the quality of something by comparing it with something else of an accepted standard:

- *rigorous benchmarking of research performance*
- *a benchmarking system to help consumers*

See

benchmark

Dictionary.Cambridge.org, 30/06/2026

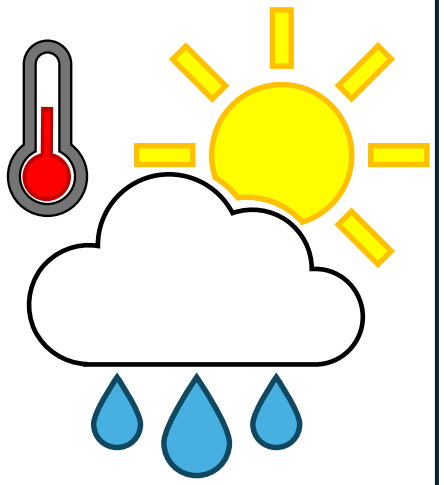
measuring the quality

something else

accepted standard

Our super amazing new model

Meteorological Observations



Met Forcing

L-A Interactions

Validation

Calibration

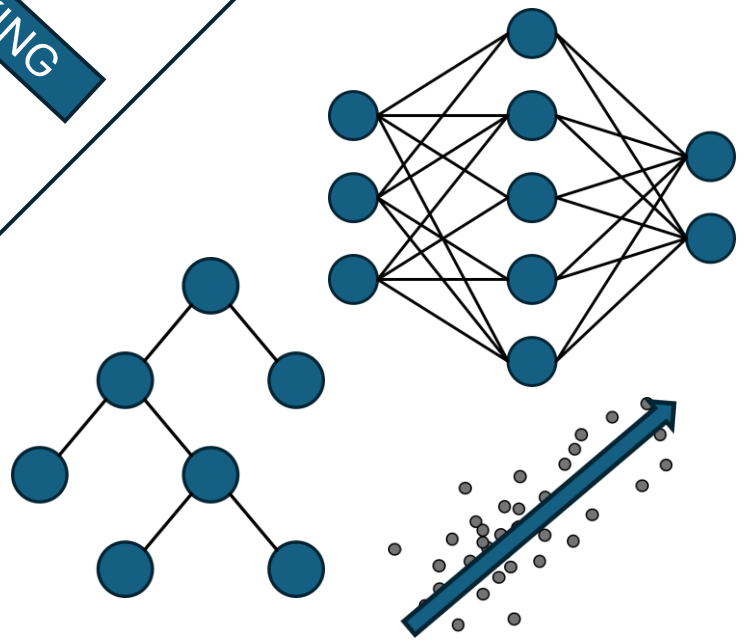
Intercomparison

BENCHMARKING

Another model

Flux Observations

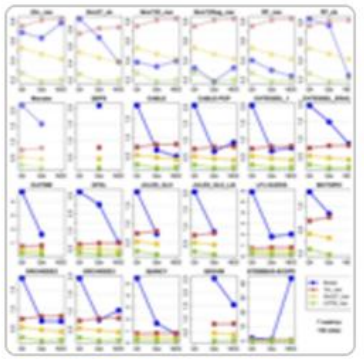
Empirical models



PLUMBER2

https://doi.org/10.5194/bg-21-5517-2024

Research article | 12 Dec 2024



On the predictability of turbulent fluxes from land: PLUMBER2 MIP experimental description and preliminary results

Gab Abramowitz ✉, Anna Ukkola, Sanaa Hobeichi, Jon Cranko Page, Mathew Lipson, Martin G. De Kauwe, Samuel Green, Claire Brenner, Jonathan Frame, Grey Nearing, Martyn Clark, Martin Best, Peter Anthoni, Gabriele Arduini, Souhail Boussetta, Silvia Caldararu, Kyeungwoo Cho, Matthias Cuntz, David Fairbairn, Craig R. Ferguson, Hyungjun Kim, Yeonjoo Kim, Jürgen Knauer, David Lawrence, Xiangzhong Luo, Sergey Malyshev, Tomoko Nitta, Jerome Ogee, Keith Oleson, Catherine Ottlé, Phillipe Peylin, Patricia de Rosnay, Heather Rumbold, Bob Su, Nicolas Vuichard, Anthony P. Walker, Xiaoni Wang-Faivre, Yunfei Wang, and Yijian Zeng

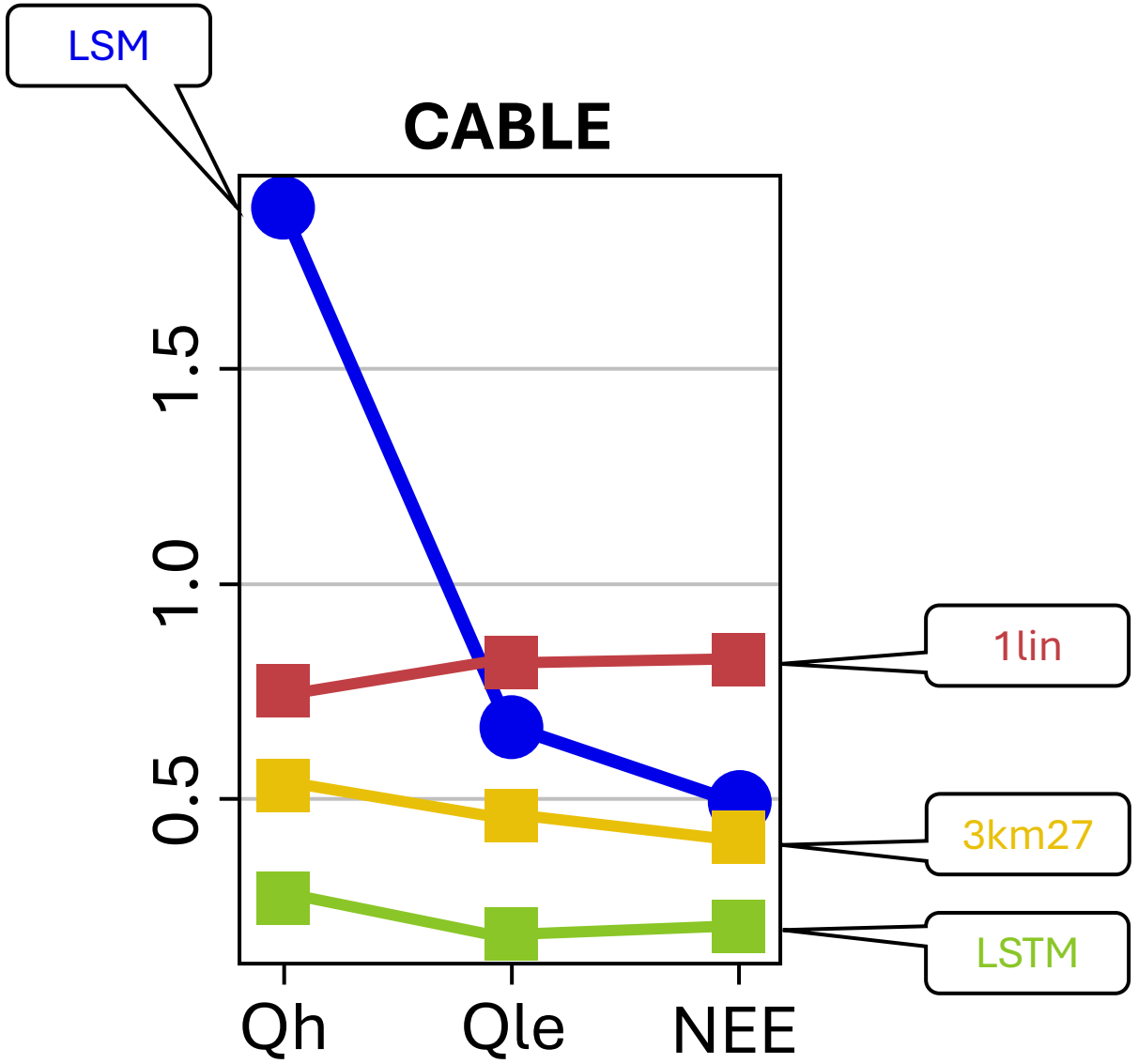
PLUMBER2

- **20 land models** simulating NEE, LE, and H as measured at **154 flux tower sites**
- Half-hourly meteorological inputs
- Land models had **no calibration to the site data**
  - Veg type, ref. height, canopy height only

- 7 (+2) empirical models as **benchmarks**
  - Linear regression, cluster + regression, Random Forest, LSTM
  - All **out-of-sample** for each site

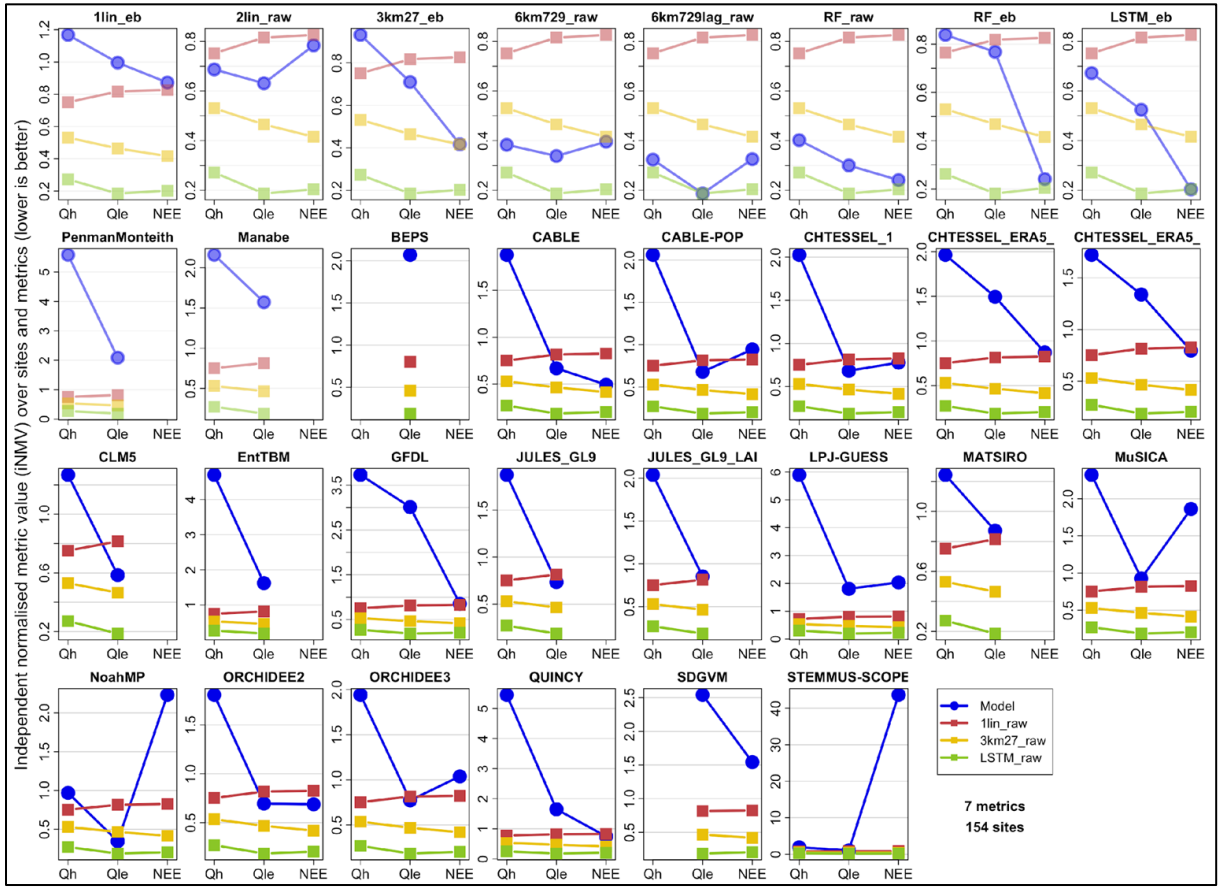
PLUMBER2

- Land models benchmarked against **hierarchy** of empirical models
- 7 metrics **normalised** to empirical model range and averaged across metrics and sites
- Land models frequently **outperformed by linear regression** against SWIN



PLUMBER2

- Land models benchmarked against **hierarchy** of empirical models
- 7 metrics **normalised** to empirical model range and averaged across metrics and sites
- Land models frequently **outperformed by linear regression** against SWIN



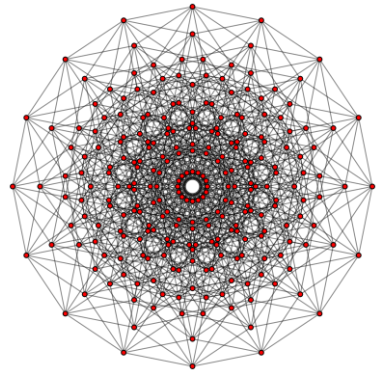
Why?

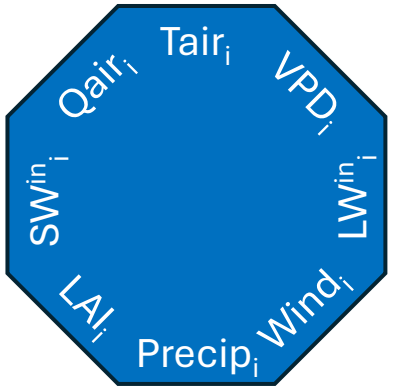
- We could look the LSM performance at each site but that's a lot of sites
- Original PLUMBER2 paper briefly considered vegetation types
  - But are these really suitable?

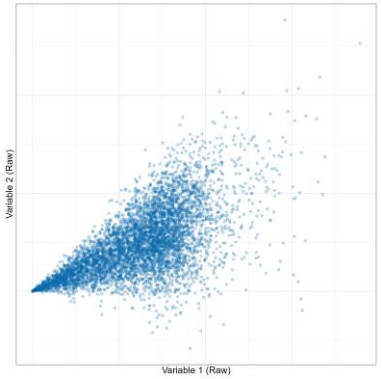
- Are there certain **meteorological conditions** where the land models are underperforming?

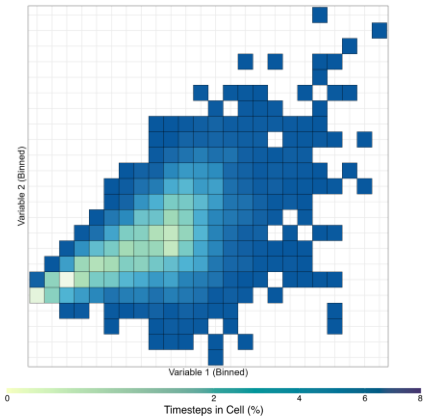
- The PLUMBER2 forcing data has 7 key met variables plus LAI
- We divide each variable into 50 equal-width bins
- This places every timestep within an 8d “cell”
- Within each cell, input conditions are similar

- We can collapse these into 2d “fingerprints” of each pair of variables







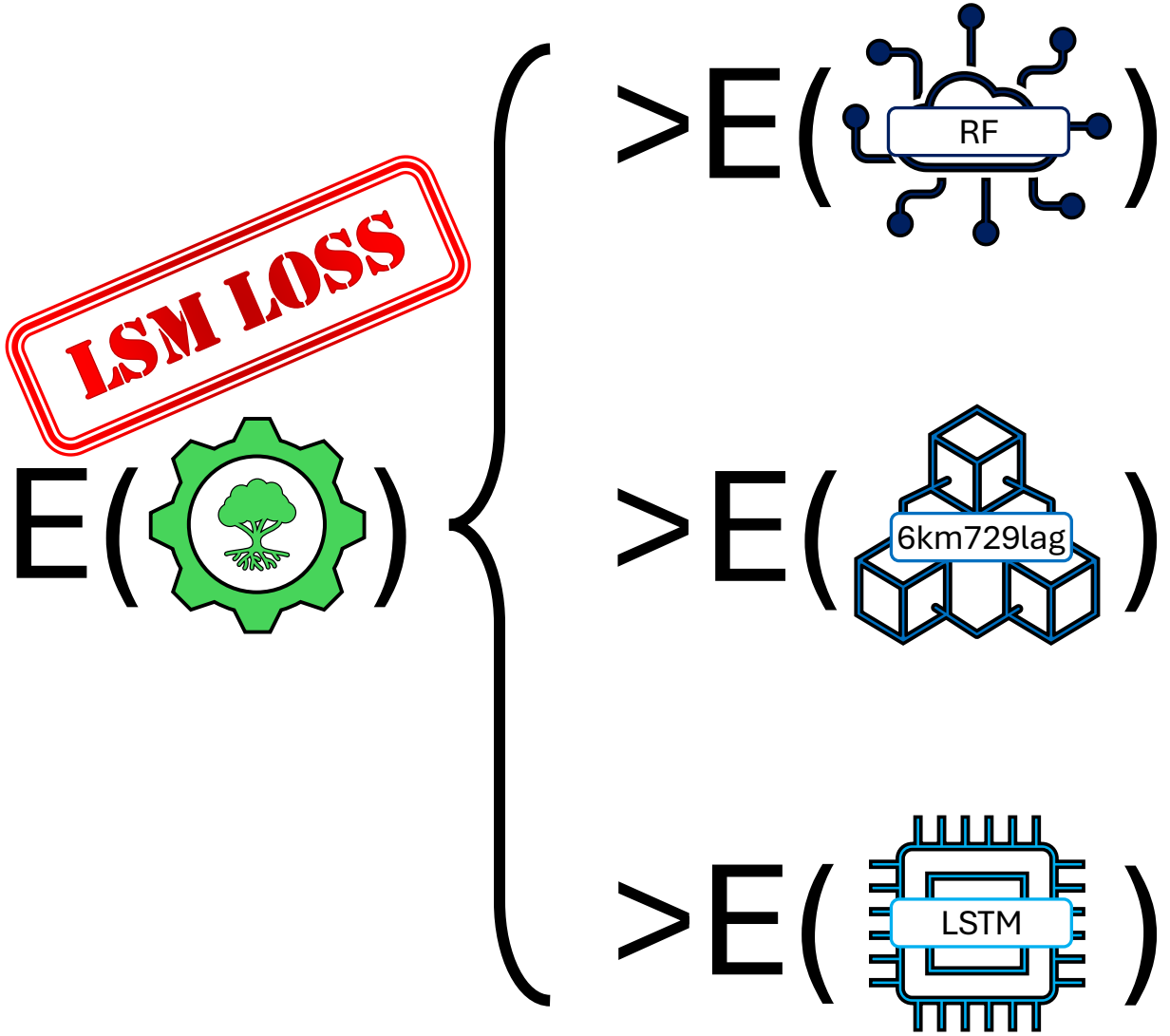


Empirical Flux Models (EFMs)

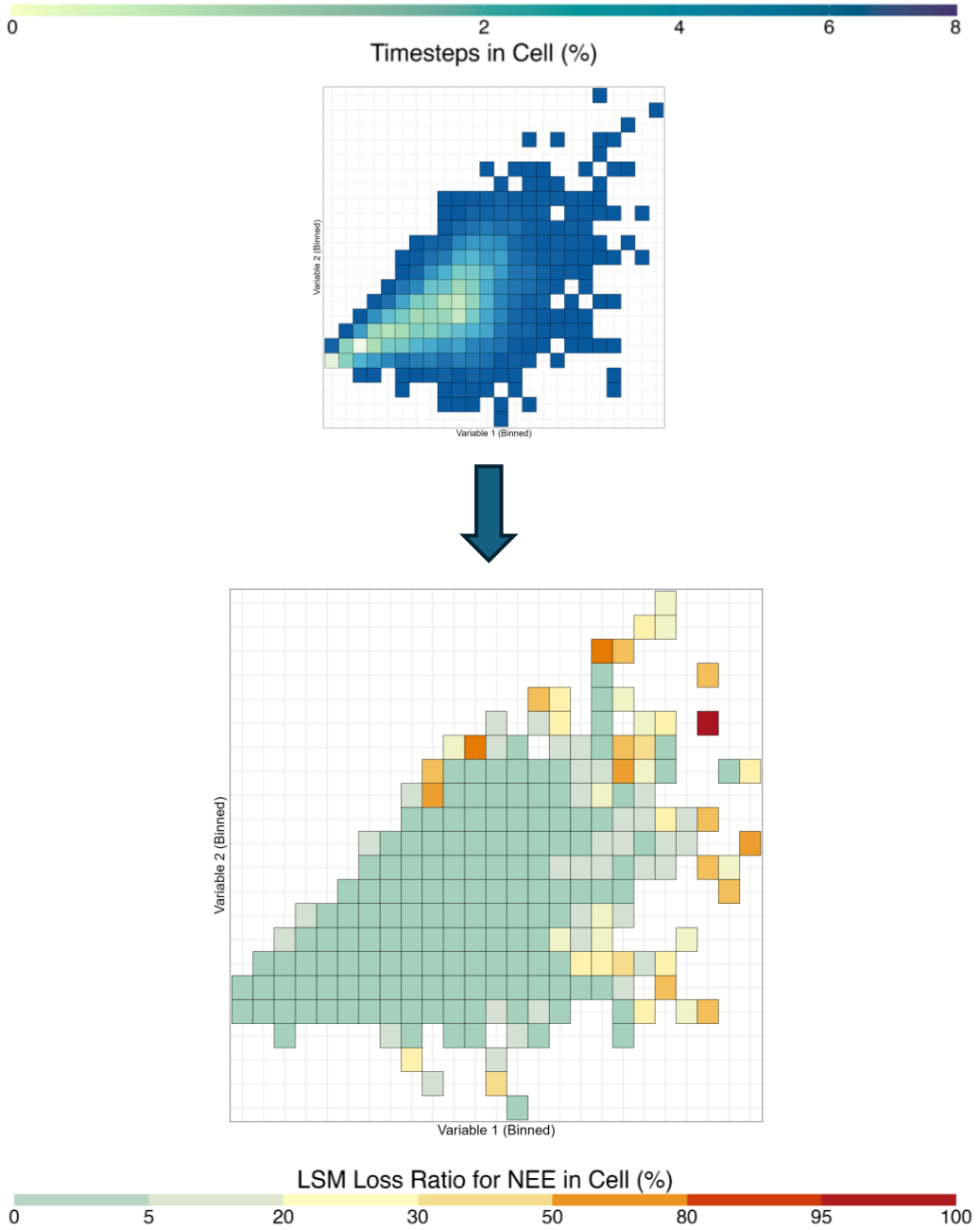
- PLUMBER2 has 3 “best” empirical models
  - 6km729lag
  - RF
  - LSTM

- At each timestep, the LSM has enough information to perform better if all 3 of these EFMs have smaller absolute error

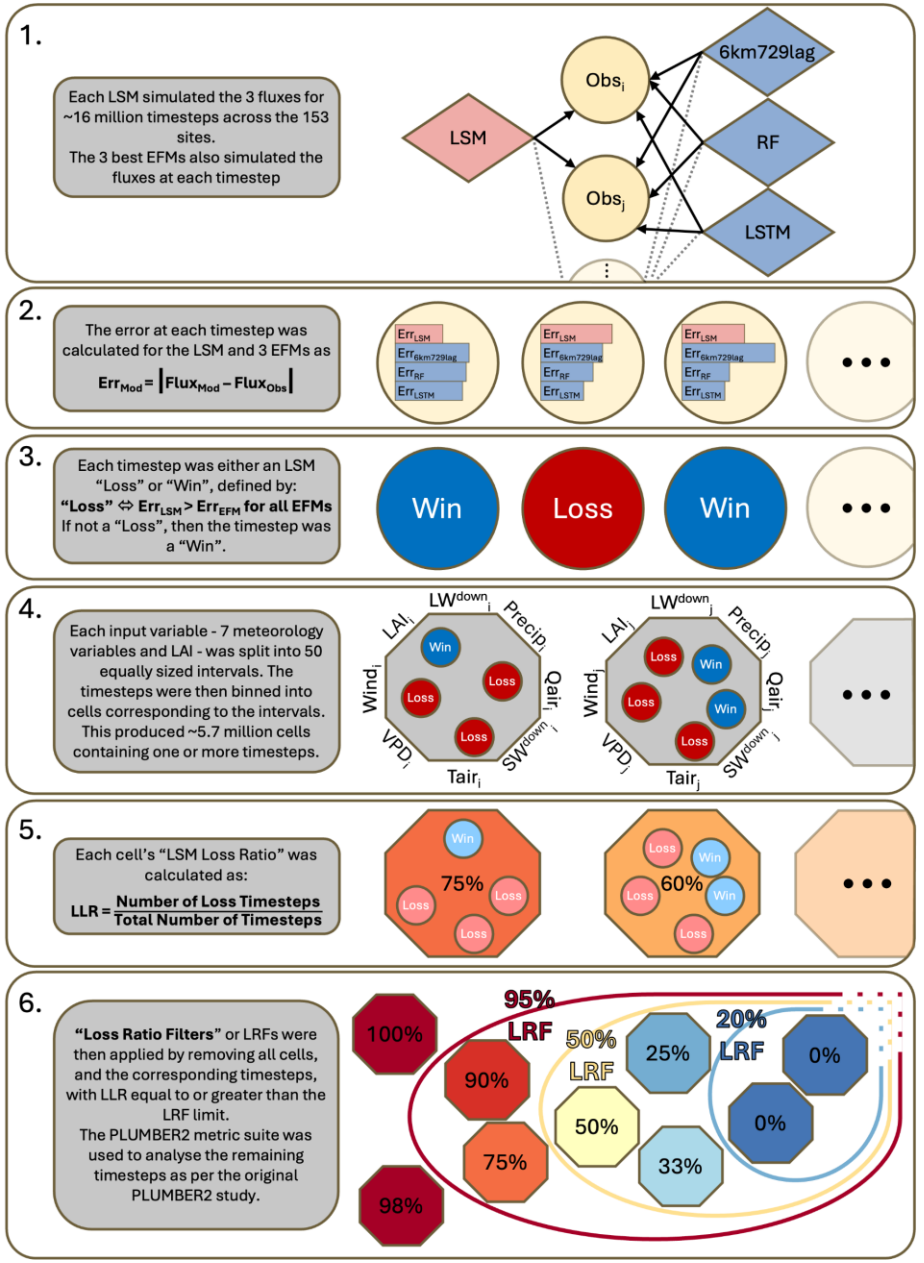
- If so, the timestep is called an “LSM Loss”



- Within each cell, we can calculate the “LSM Loss Ratio”
  - What percentage of timesteps within the cell are an LSM Loss?
  - Null expectation (under random uncertainty) is 25%
  - Over 25% implies our LSMs are underperforming

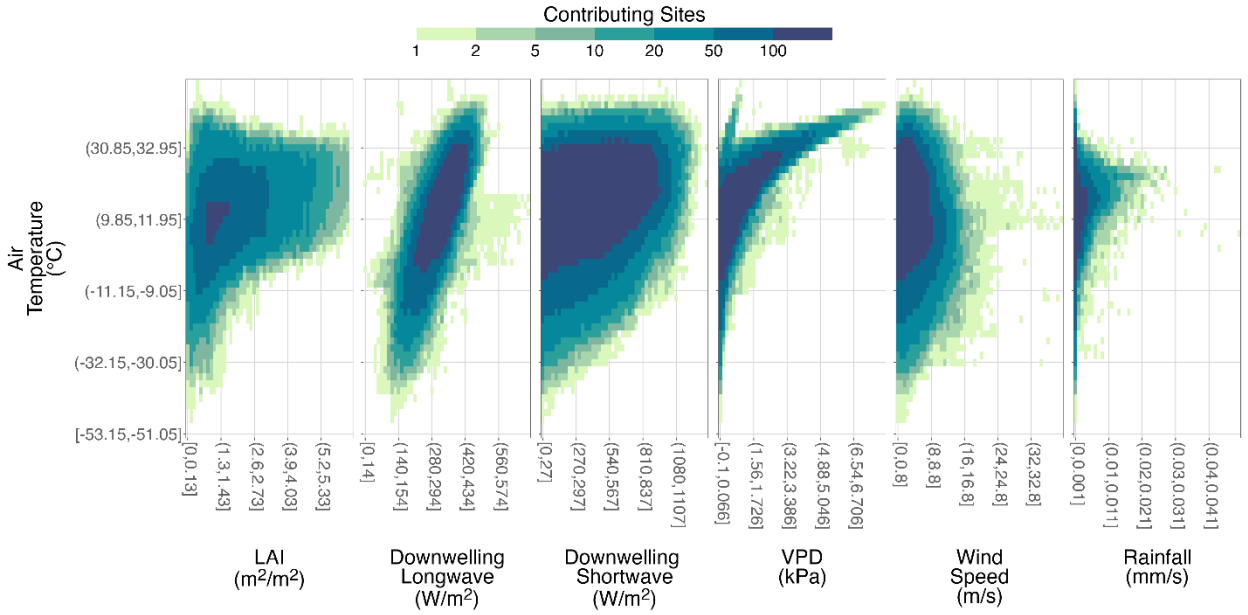


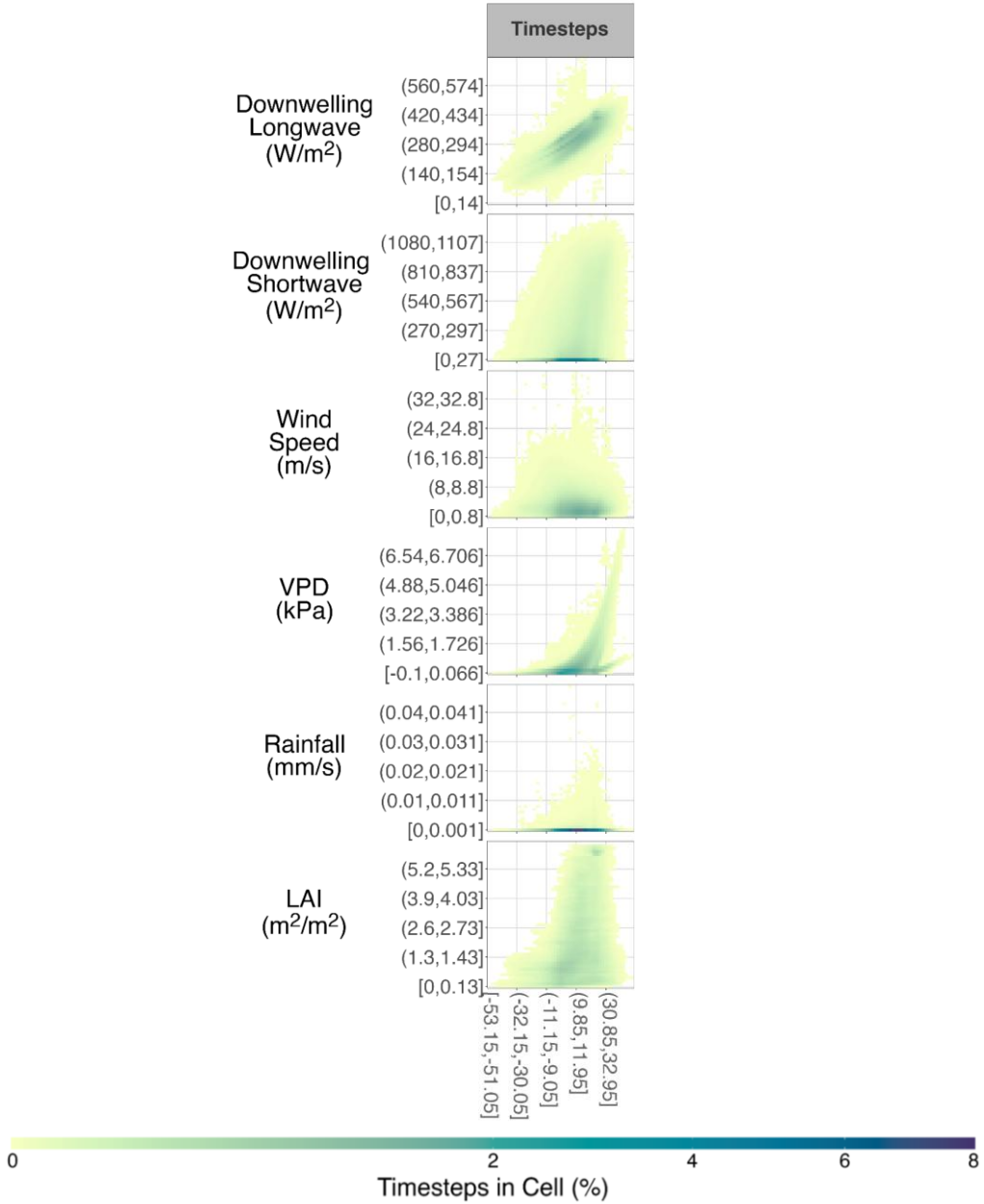
- The manuscript has a schematic to explain this set up in more detail



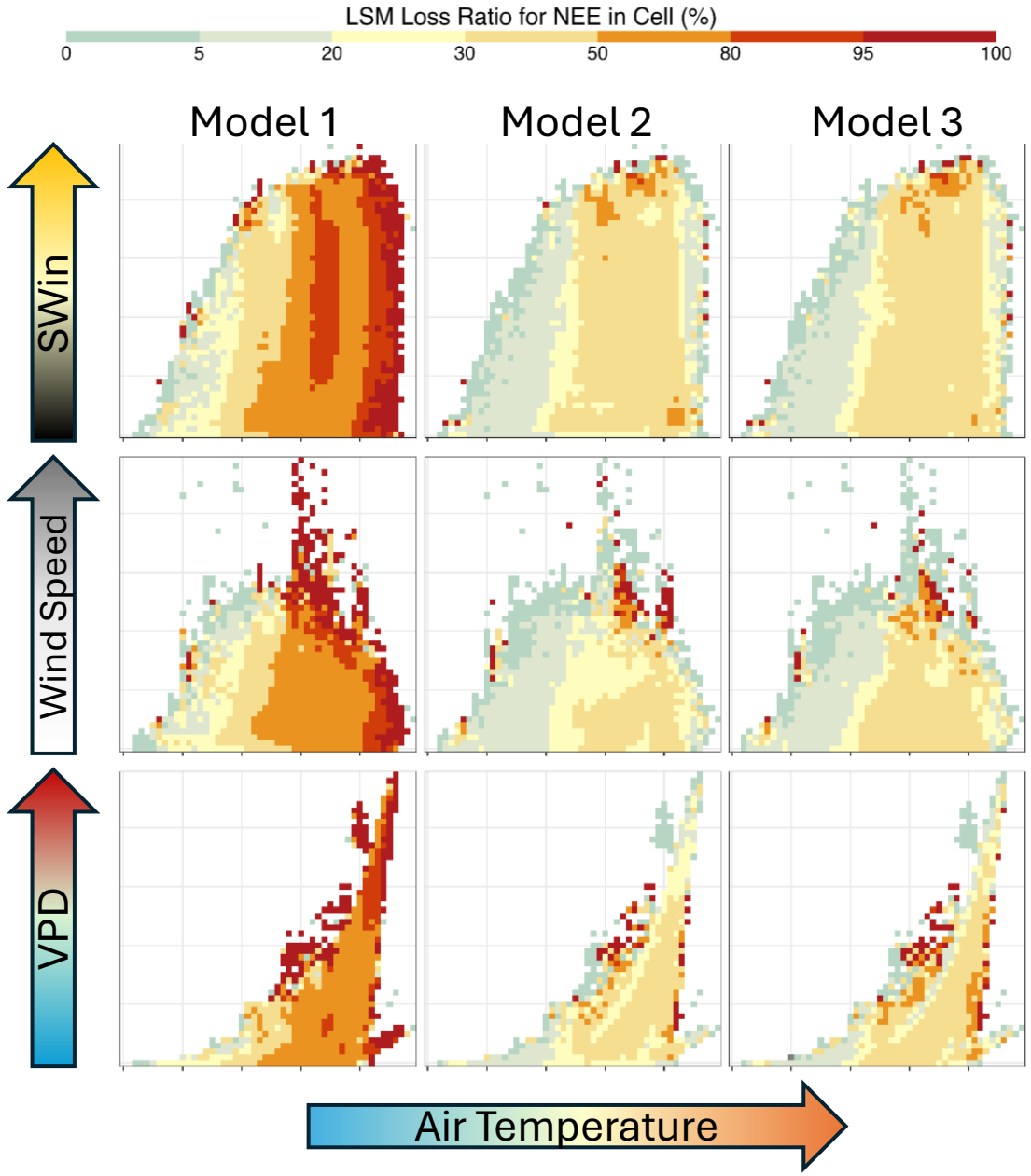
- What do our 2d met domains look like?

- Edge cases have fewer timesteps from fewer sites

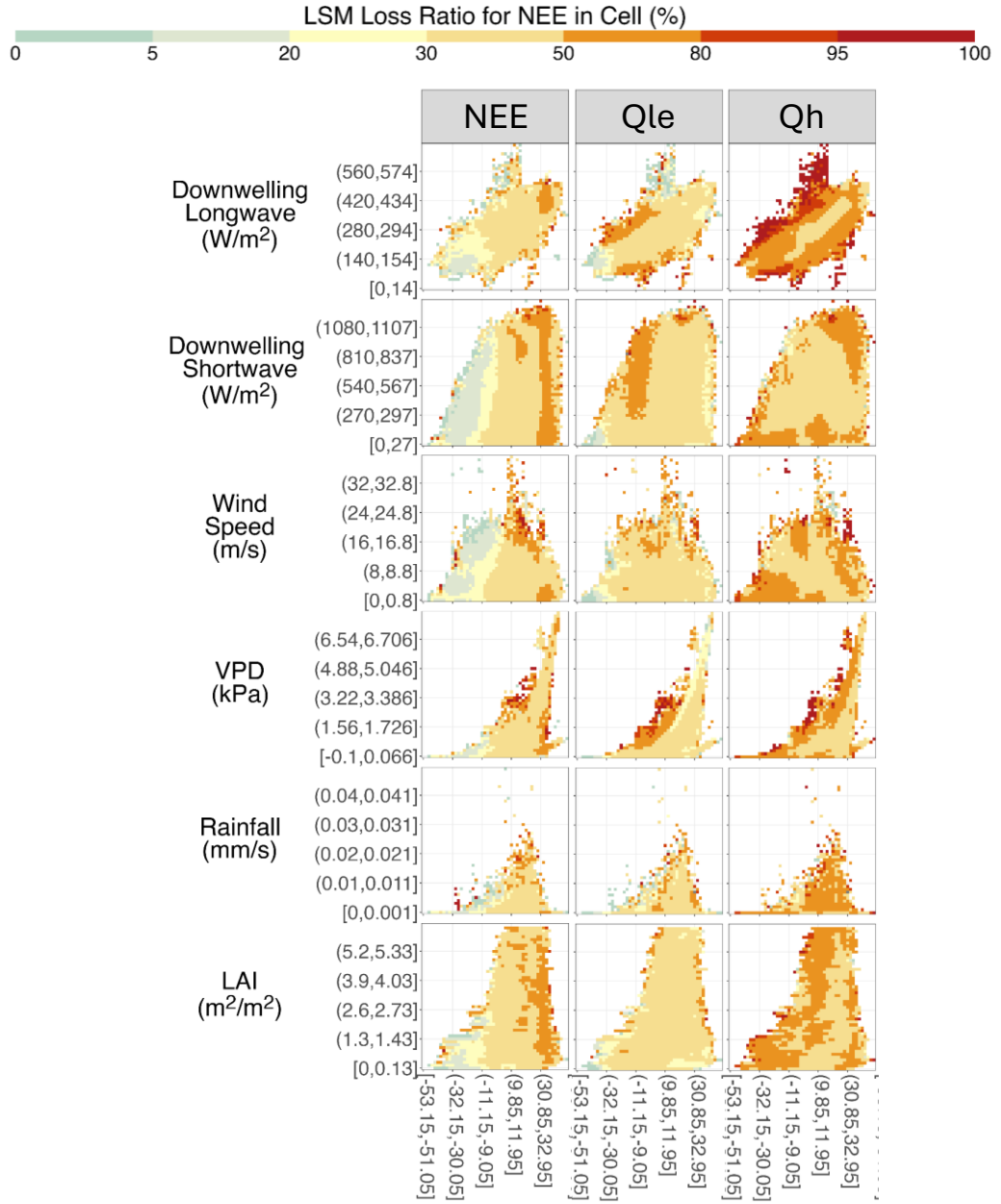


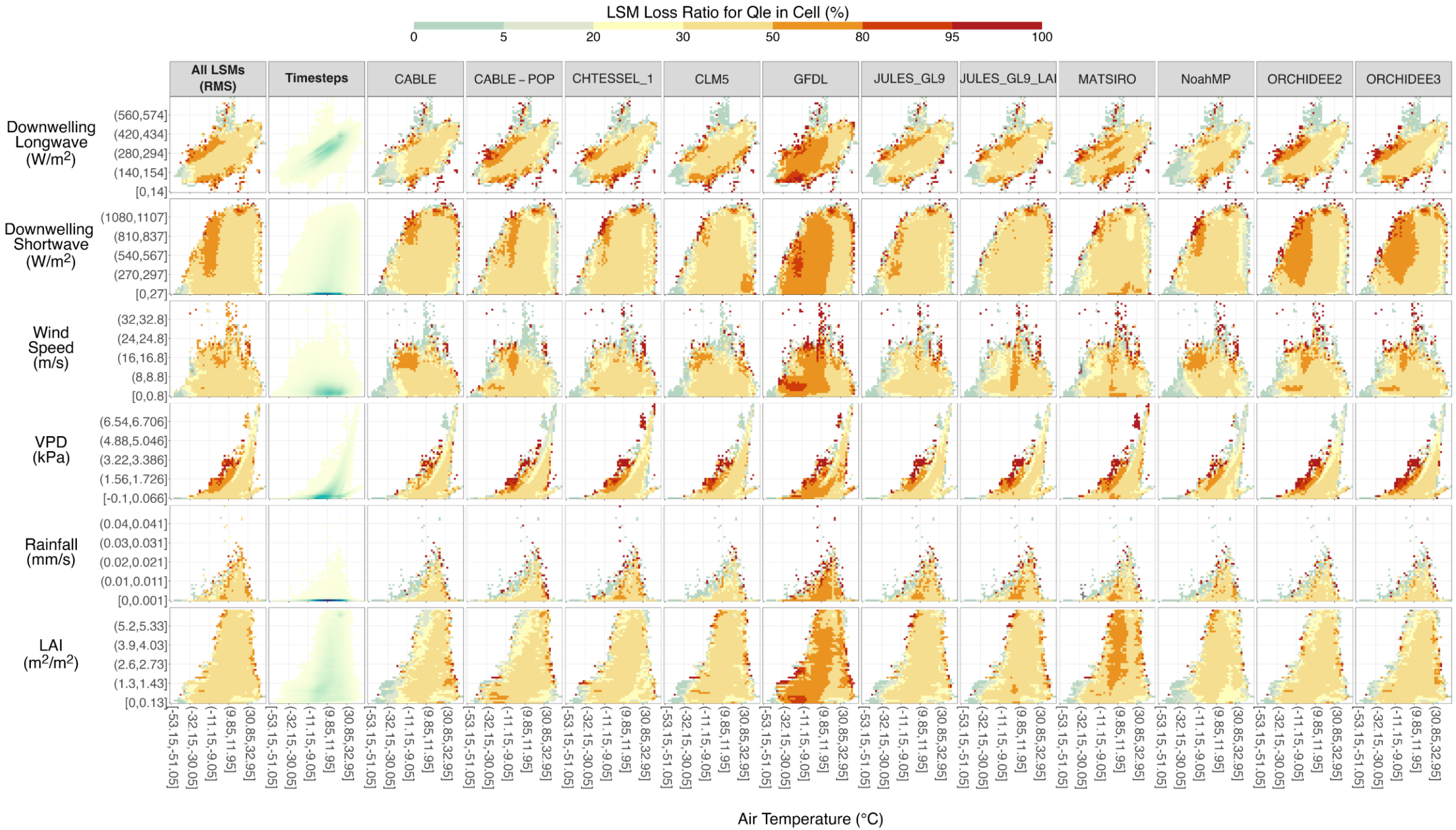


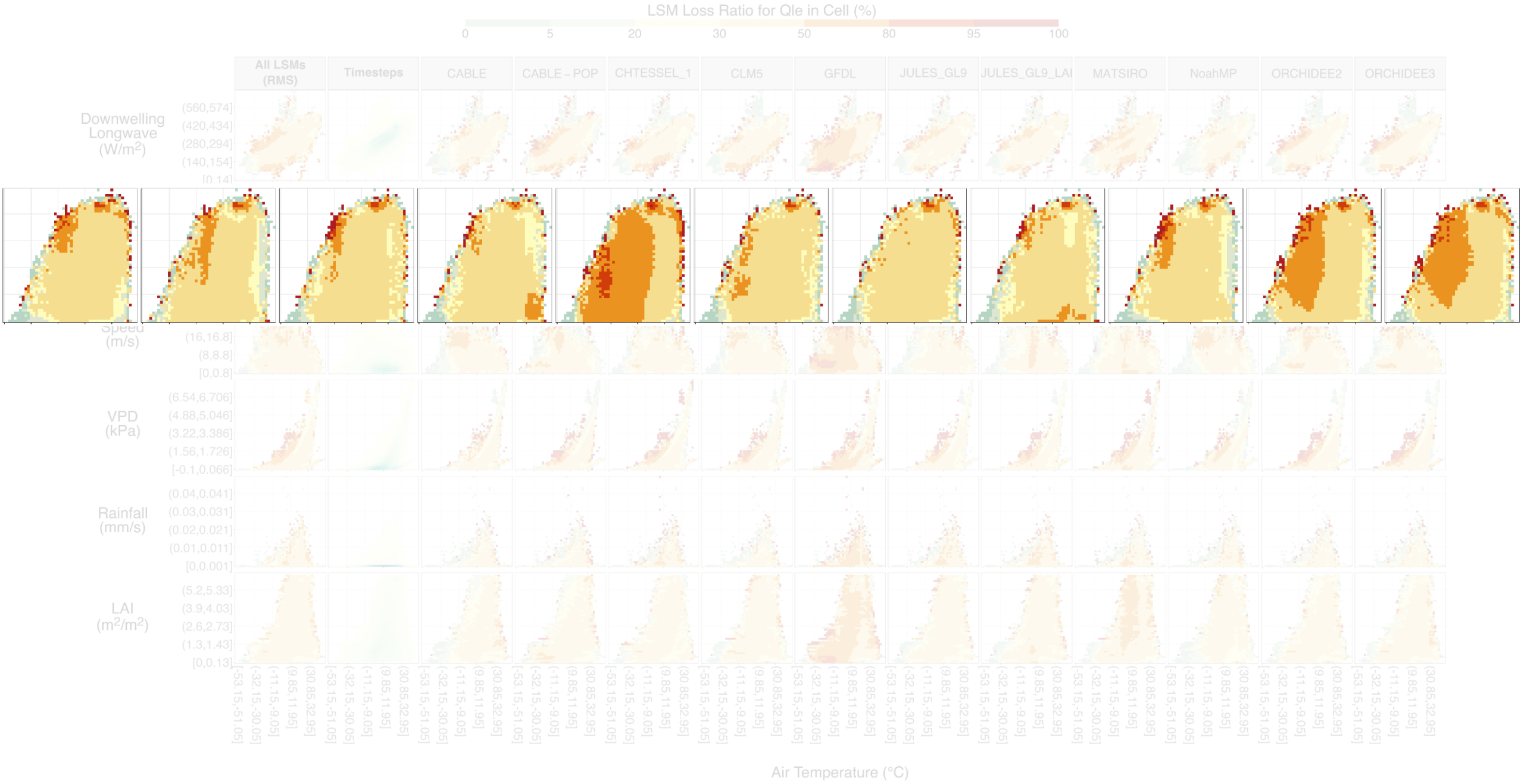
- LSM Loss Ratio for NEE from 3 LSMs
- All 3 perform well at low temperatures
- Model 1 performance has strong temperature dependence
- Models 2 and 3 are same family
- The highest LLR occur in edge cells

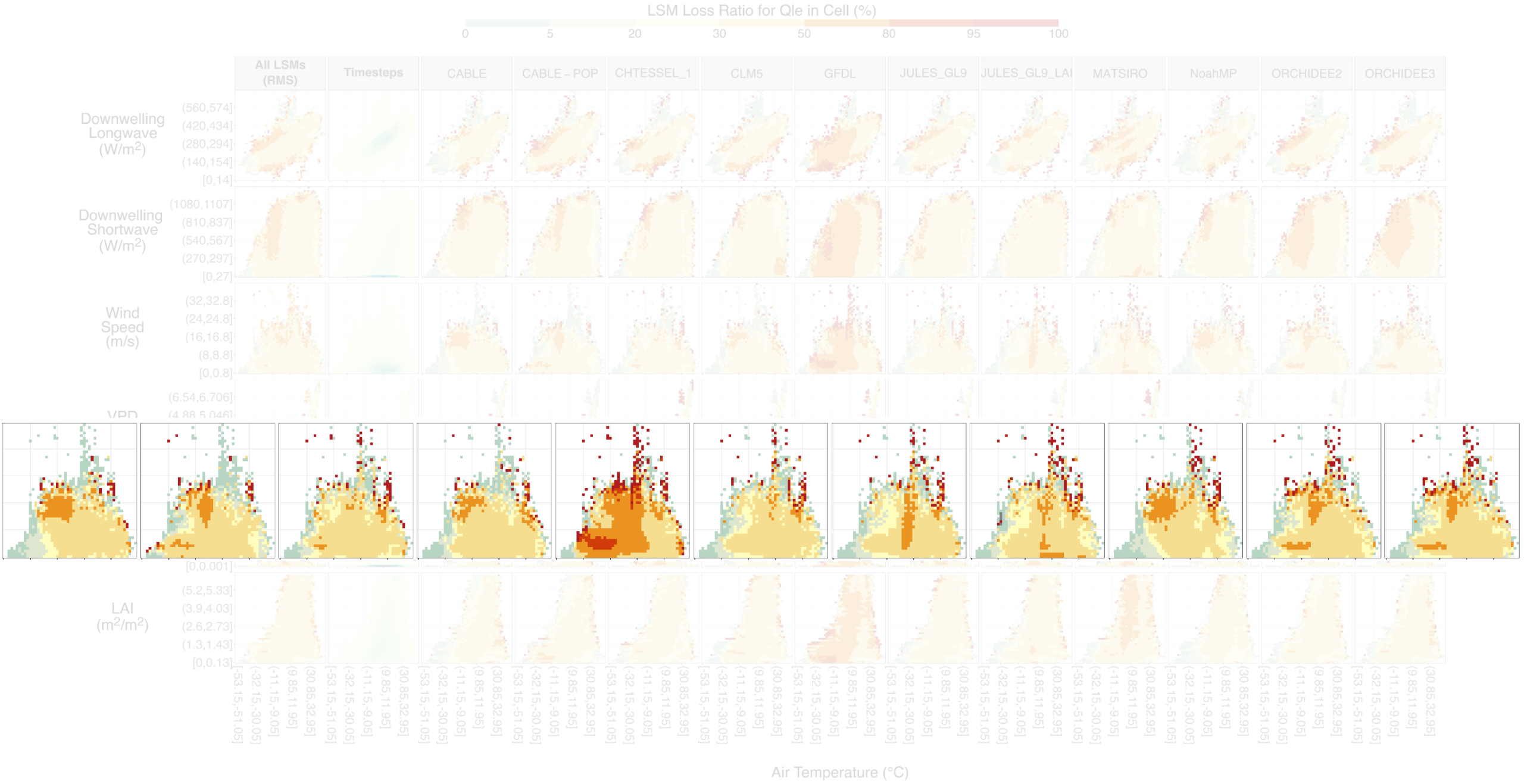


- LSM Loss Ratio across all 11 LSMs in the study
- Root mean square error
- Three fluxes
  - NEE = carbon
  - Qle = latent heat
  - Qh = sensible heat
- Qh noticeably worse as expected from PLUMBER2




LSM Loss Ratio for Qle in Cell (%)
0
5
20
30
50
80
95
100
All LSMs (RMS)
Timesteps
CABLE
CABLE – POP
CHTESSEL_1
CLM5
GFDL
JULES_GL9
JULES_GL9_LAI
MATSIRO
NoahMP
ORCHIDEE2
ORCHIDEE3
Downwelling Longwave (W/m2)
(560,574]
(420,434]
(280,294]
(140,154]
[0,14]
Downwelling Shortwave (W/m2)
(1080,1107]
(810,837]
(540,567]
(270,297]
[0,27]
Wind Speed (m/s)
(32,32.8]
(24,24.8]
(16,16.8]
(8,8.8]
[0,0.8]
VPD (kPa)
(6.54,6.706]
(4.88,5.046]
(3.22,3.386]
(1.56,1.726]
[-0.1,0.066]
Rainfall (mm/s)
(0.04,0.041]
(0.03,0.031]
(0.02,0.021]
(0.01,0.011]
[0,0.001]
LAI (m2/m2)
(5.2,5.33]
(3.9,4.03]
(2.6,2.73]
(1.3,1.43]
[0,0.13]
[-53.15,-51.05]
(-32.15,-30.05]
(-11.15,-9.05]
(9.85,11.95]
(30.85,32.95]
Air Temperature (°C)

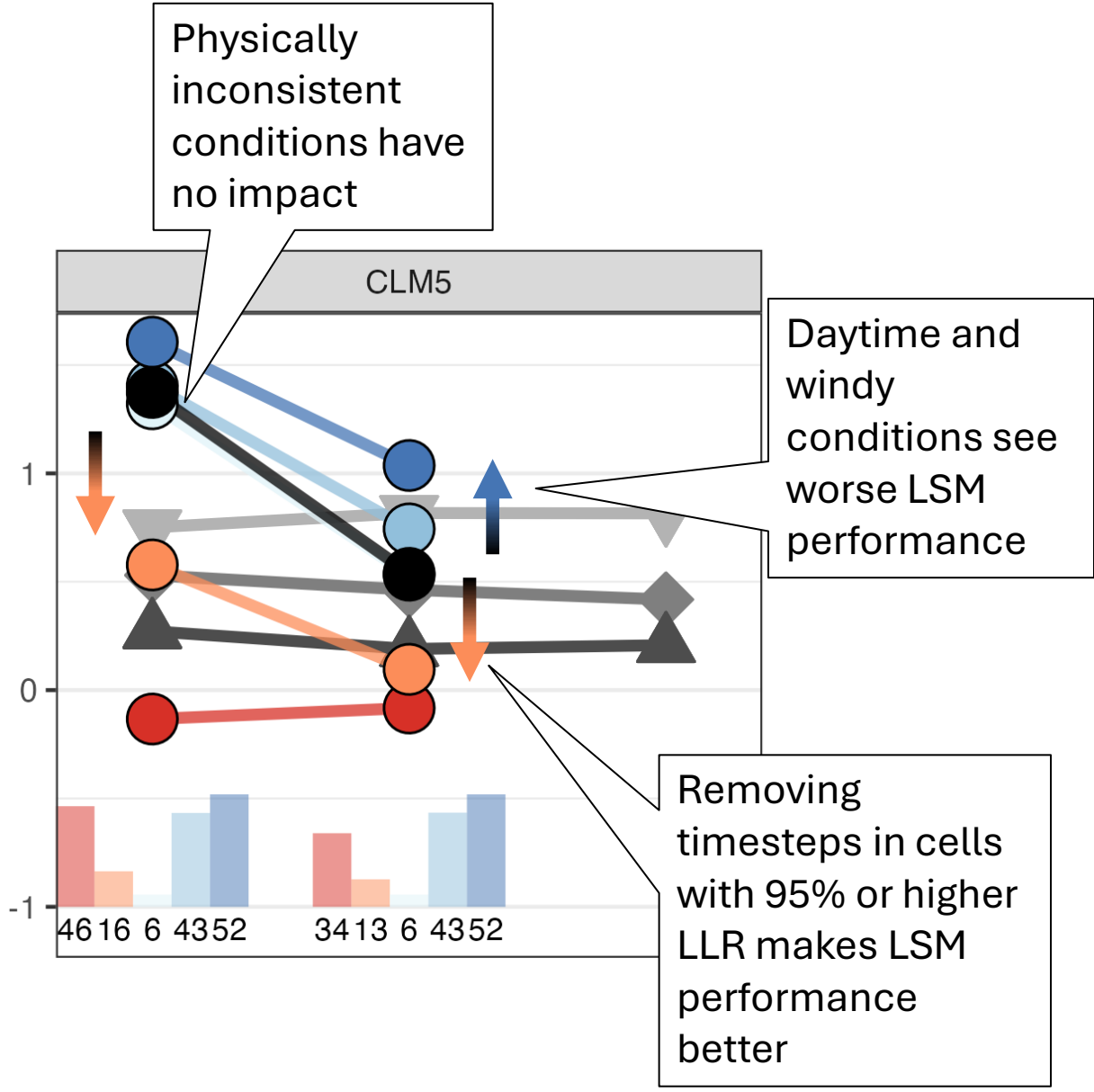

LSM Loss Ratio for Qle in Cell (%)
0
5
20
30
50
80
95
100
All LSMs (RMS)
Timesteps
CABLE
CABLE – POP
CHTESSEL_1
CLM5
GFDL
JULES_GL9
JULES_GL9_LAI
MATSIRO
NoahMP
ORCHIDEE2
ORCHIDEE3
Downwelling Longwave (W/m²)
(560,574]
(420,434]
(280,294]
(140,154]
[0,14]
Speed (m/s)
(16,16.8]
(8,8.8]
[0,0.8]
VPD (kPa)
(6.54,6.706]
(4.88,5.046]
(3.22,3.386]
(1.56,1.726]
[-0.1,0.066]
Rainfall (mm/s)
(0.04,0.041]
(0.03,0.031]
(0.02,0.021]
(0.01,0.011]
[0,0.001]
LAI (m²/m²)
(5.2,5.33]
(3.9,4.03]
(2.6,2.73]
(1.3,1.43]
[0,0.13]
[-53.15,-51.05]
(-32.15,-30.05]
(-11.15,-9.05]
(9.85,11.95]
(30.85,32.95]
Air Temperature (°C)


LSM Loss Ratio for Qle in Cell (%)
0
5
20
30
50
80
95
100
All LSMs (RMS)
Timesteps
CABLE
CABLE – POP
CHTESSEL_1
CLM5
GFDL
JULES_GL9
JULES_GL9_LAI
MATSIRO
NoahMP
ORCHIDEE2
ORCHIDEE3
Downwelling Longwave (W/m²)
(560,574]
(420,434]
(280,294]
(140,154]
[0,14]
Downwelling Shortwave (W/m²)
(1080,1107]
(810,837]
(540,567]
(270,297]
[0,27]
Wind Speed (m/s)
(32,32.8]
(24,24.8]
(16,16.8]
(8,8.8]
[0,0.8]
(6.54,6.706]
[0,0.001]
LAI (m²/m²)
(5.2,5.33]
(3.9,4.03]
(2.6,2.73]
(1.3,1.43]
[0,0.13]
[-53.15,-51.05]
(-32.15,-30.05]
(-11.15,-9.05]
(9.85,11.95]
(30.85,32.95]
Air Temperature (°C)

- The timesteps within these cells are responsible for the PLUMBER2 result

- Removing them from the PLUMBER2 analysis can dramatically improve the LSMs' performance

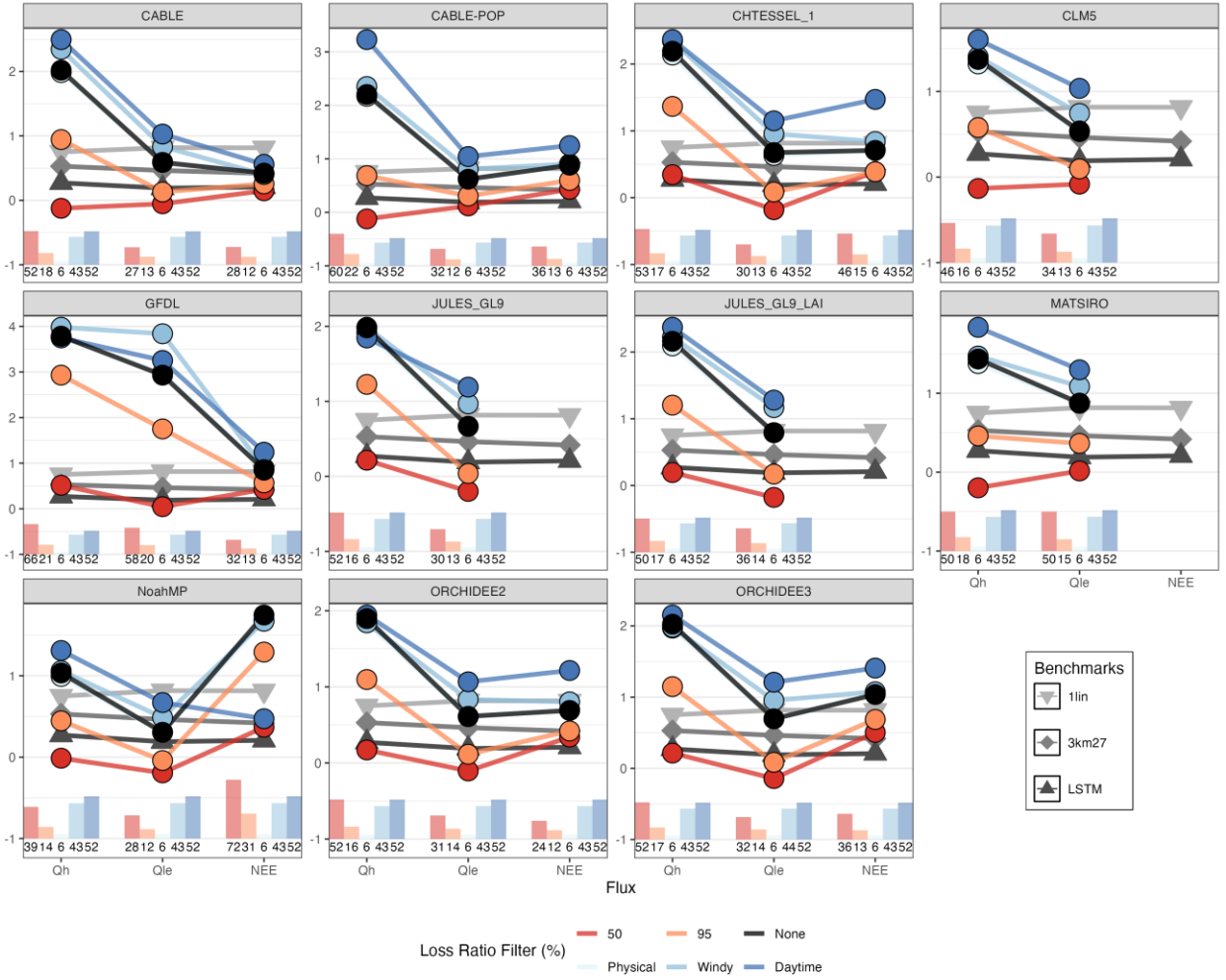
- Other types of timestep filters used
  - Daytime only (SWdown > 10 Wm^{-2})
  - Windy (Wind > 2 ms^{-1})
  - Physically plausible



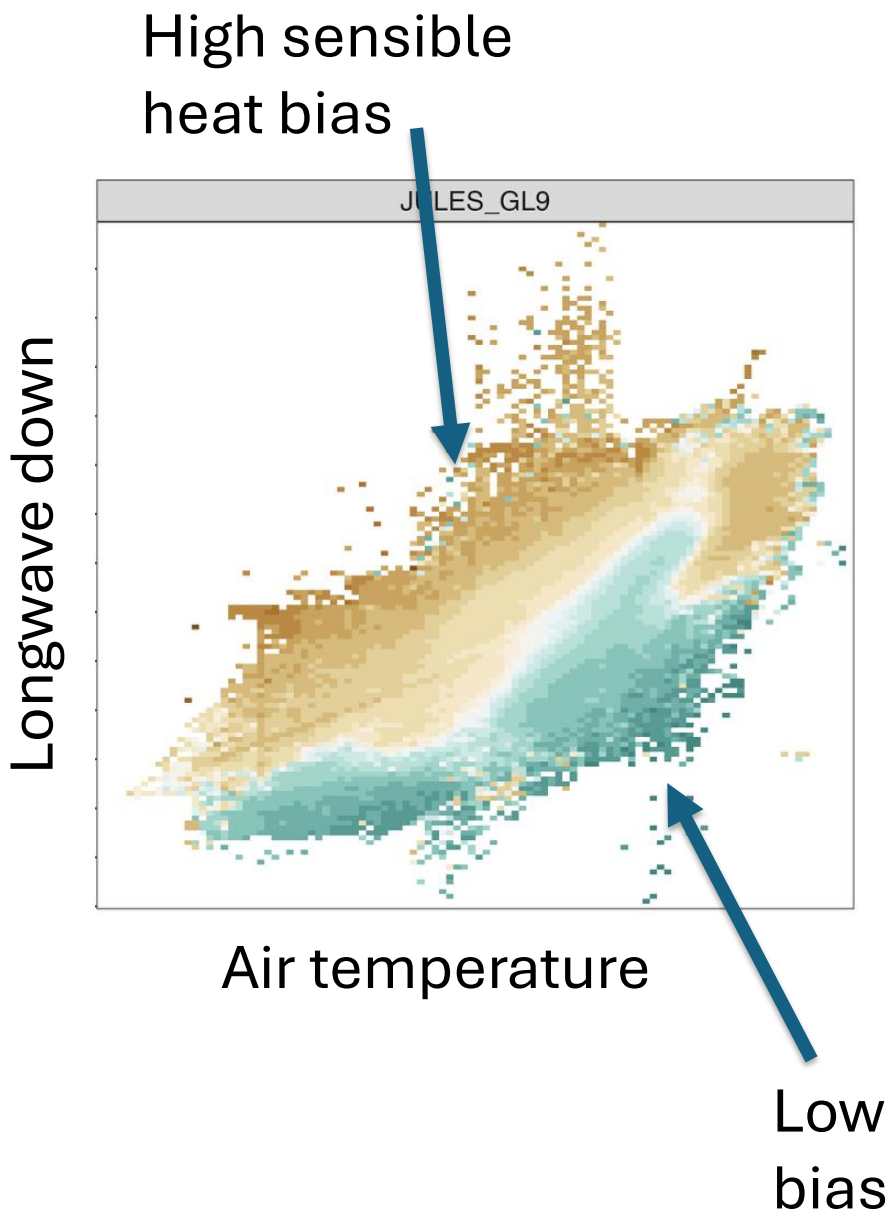
- The timesteps within these cells are responsible for the PLUMBER2 result

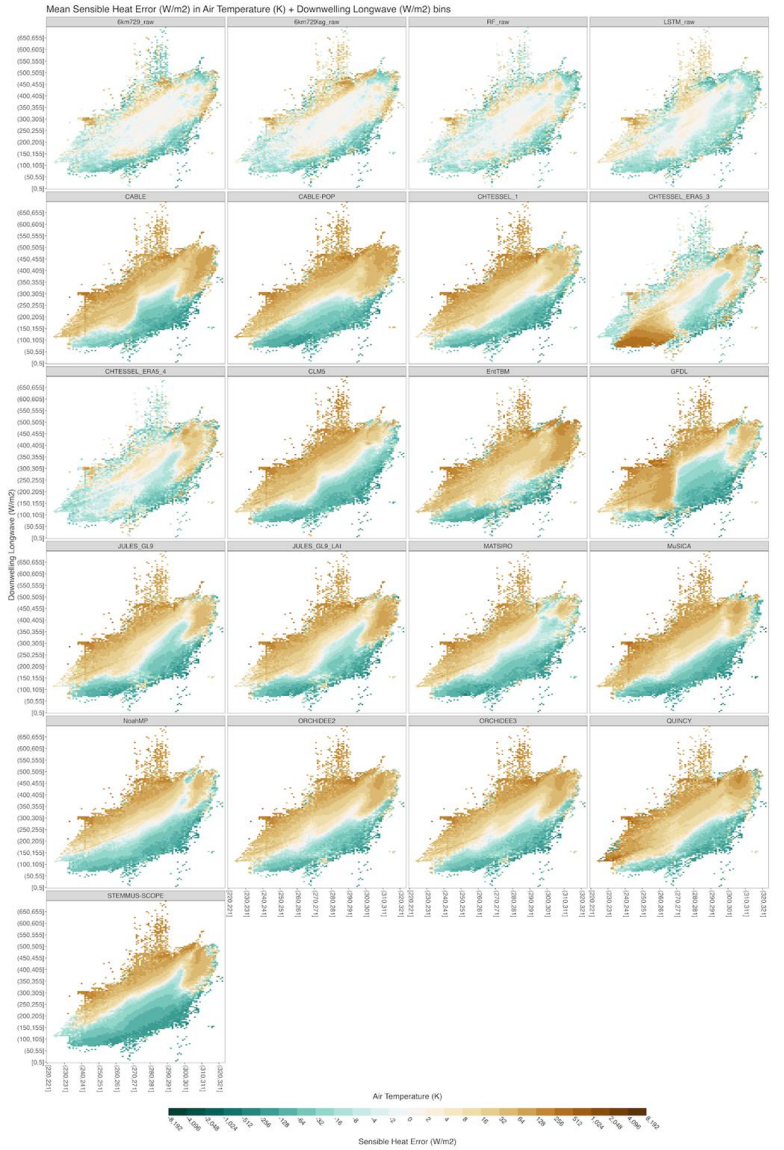
- Removing them from the PLUMBER2 analysis can dramatically improve the LSMs' performance

- Other types of timestep filters used
  - Daytime only (SWdown > 10 Wm^{-2})
  - Windy (Wind > 2 ms^{-1})
  - Physical plausible

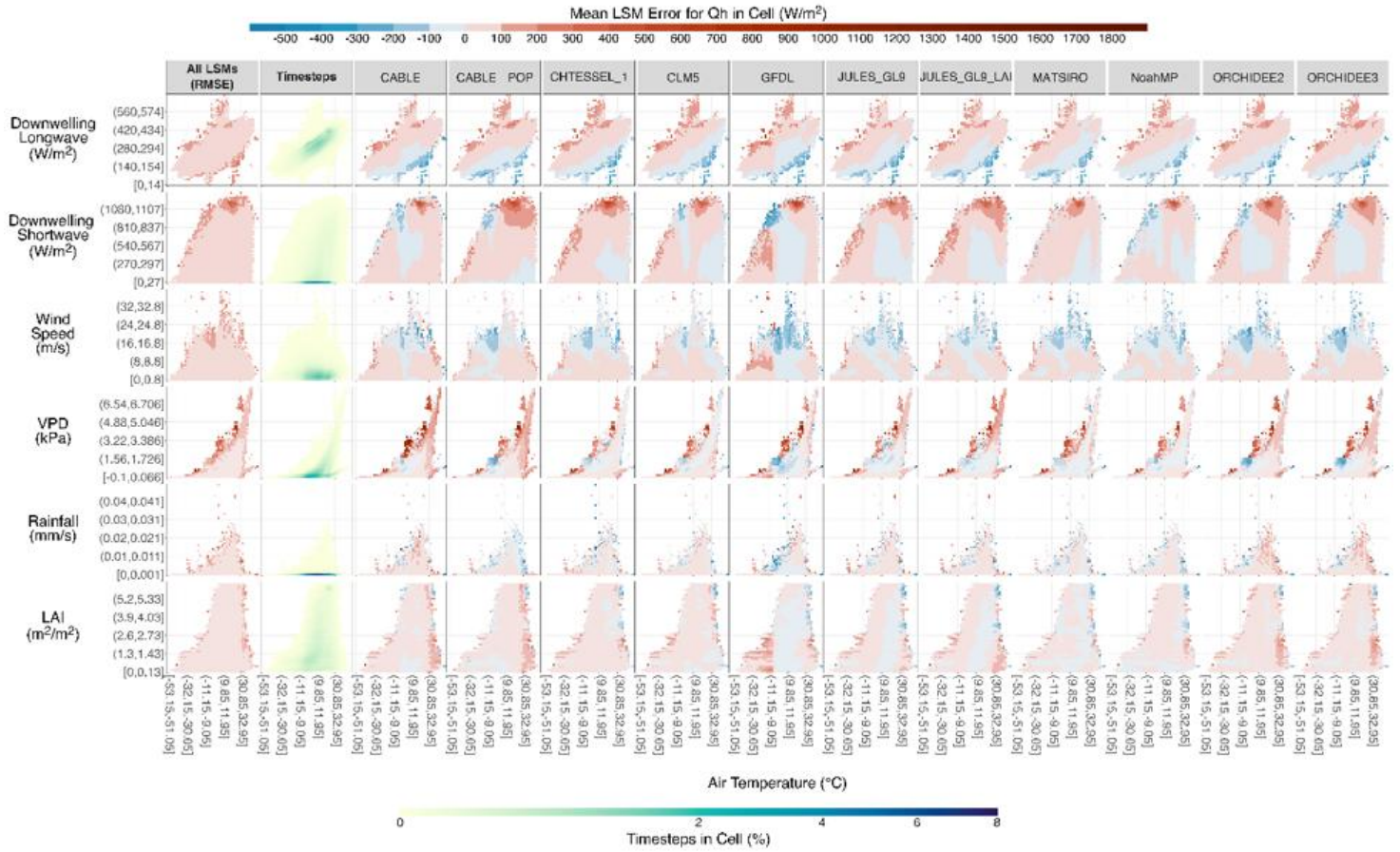


- Similar fingerprinting reveals consistent LSM bias
- Implemented in modelevaluation.org
- Strong potential for directing model development

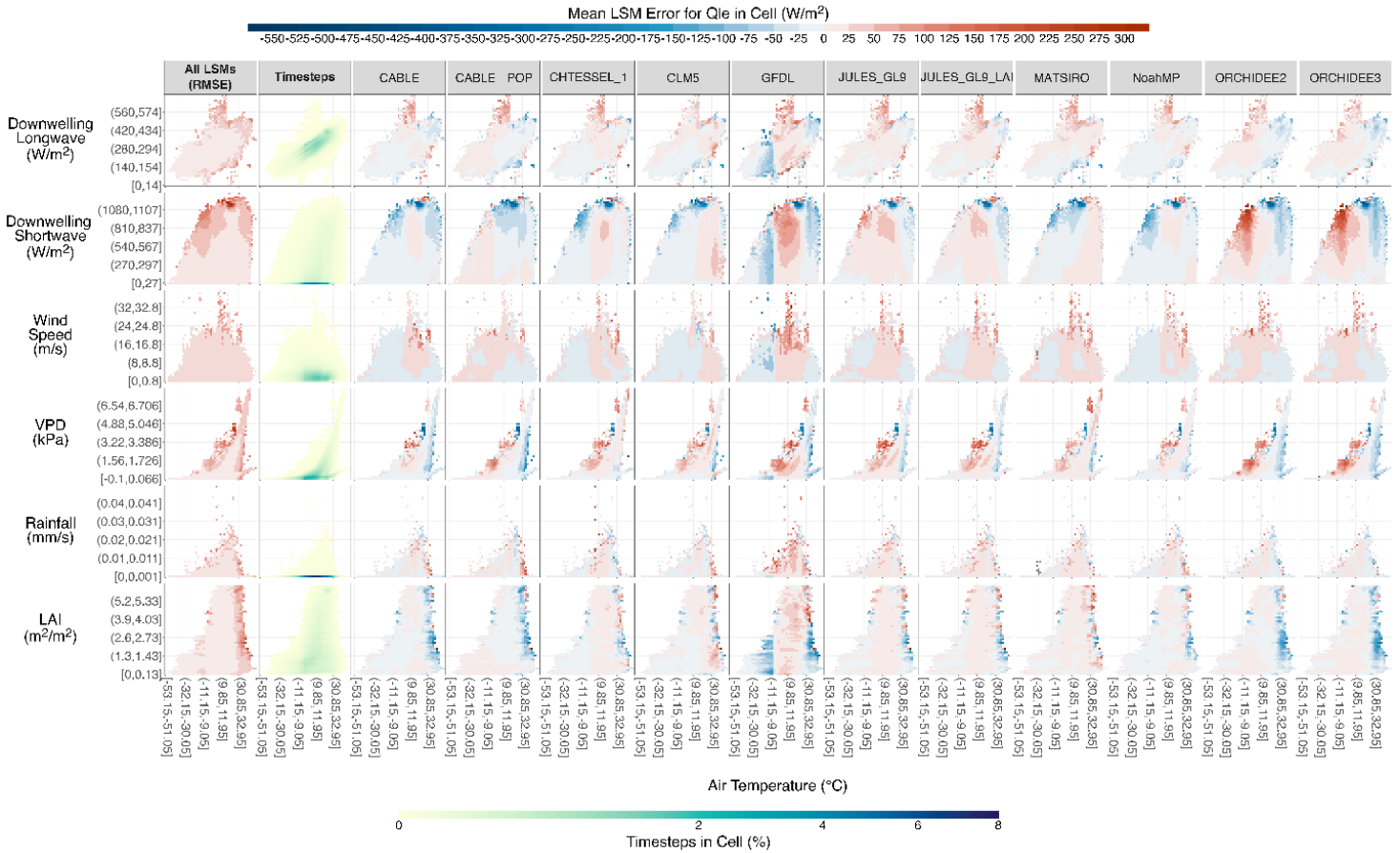




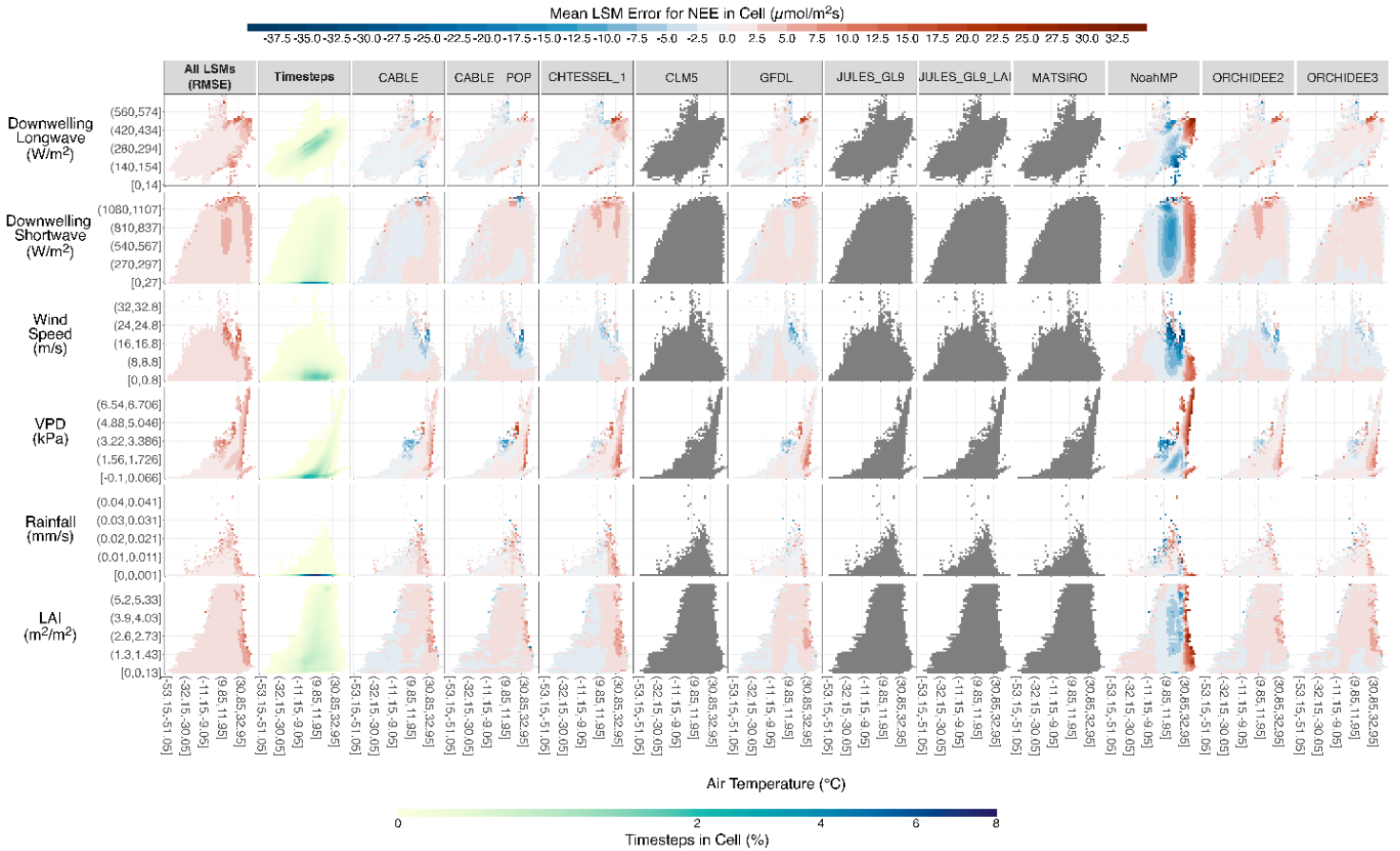
- Similar fingerprinting reveals consistent LSM bias
- Implemented in modelevaluation.org
- Strong potential for directing model development



- Similar fingerprinting reveals consistent LSM bias
- Implemented in modelevaluation.org
- Strong potential for directing model development



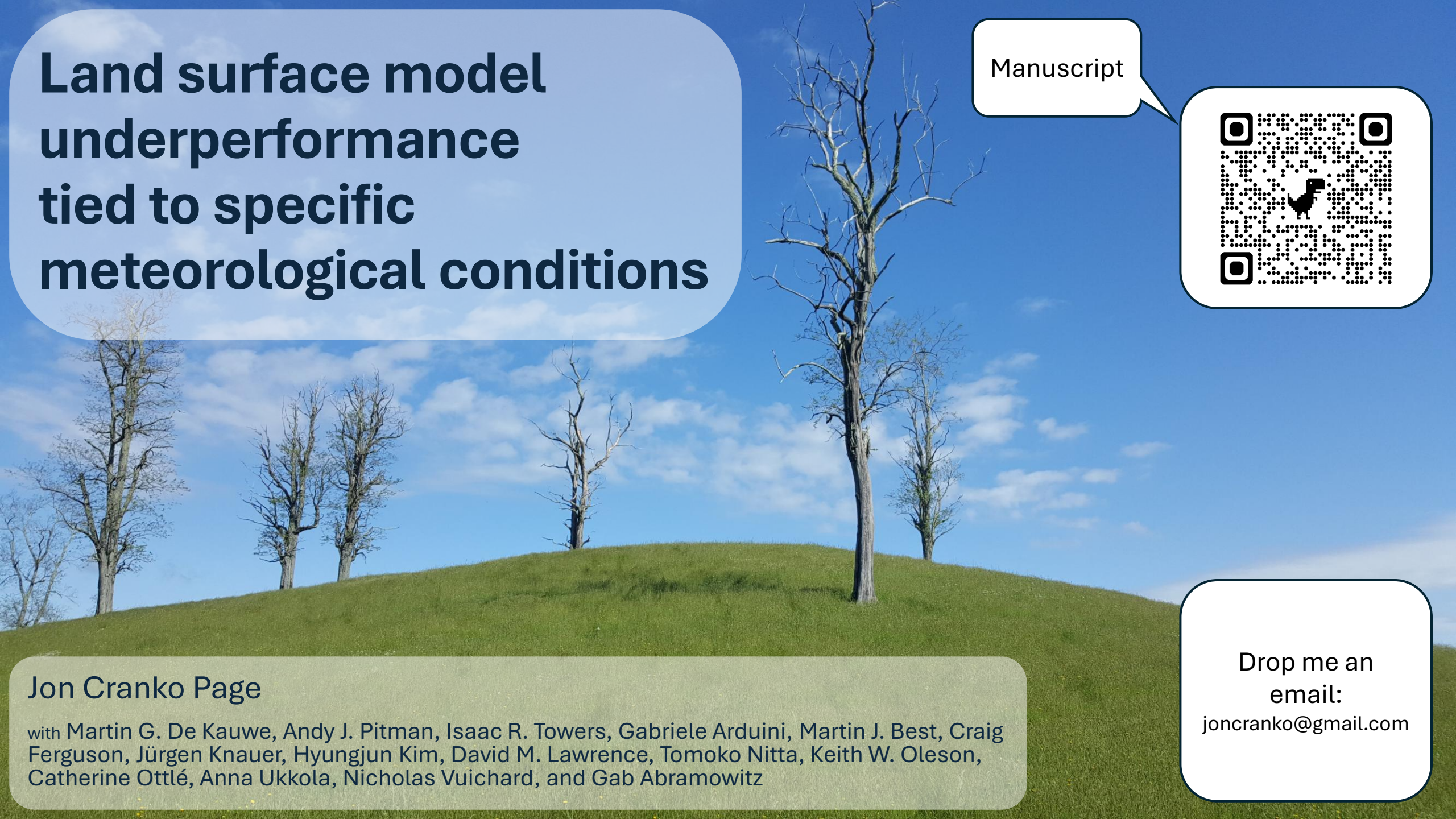
- Similar fingerprinting reveals consistent LSM bias
- Implemented in modelevaluation.org
- Strong potential for directing model development



Our study based on the **benchmarking framework PLUMBER2**
identified that a collection of **land surface models underperformed**
simulating carbon, water, and energy fluxes **relative to empirical benchmarks**
when **two (or more) meteorological drivers** were **comparatively extreme**

When **meteorological conditions** are at the **edge of the joint distribution** of the input domain space,
our **land surface models should be doing better** with the information they are provided.

Land surface model underperformance tied to specific meteorological conditions



Manuscript

Jon Cranko Page

with Martin G. De Kauwe, Andy J. Pitman, Isaac R. Towers, Gabriele Arduini, Martin J. Best, Craig Ferguson, Jürgen Knauer, Hyungjun Kim, David M. Lawrence, Tomoko Nitta, Keith W. Oleson, Catherine Ottlé, Anna Ukkola, Nicholas Vuichard, and Gab Abramowitz

Drop me an email:
joncranko@gmail.com