

Reviewer 2

Overall, I think this is a well-structured and valuable manuscript. The proposed framework provides a useful step toward observationally constrained land-surface emulation, and the manuscript is generally clear and comprehensive. I recommend minor revision, primarily to clarify several methodological choices and further discuss potential future developments.

We thank the reviewer for their careful and constructive reading of the manuscript, and for their positive assessment. We have addressed each of the four comments below and have kept additions to the minimum needed to answer the point, since the manuscript is already quite long.

The current study adopts an MLP architecture, which provides a good balance between computational efficiency and predictive performance. However, the current architecture does not explicitly represent spatial or longer-term temporal dependencies. It would be useful to expand the Discussion on potential future ML developments, such as architectures that can better capture spatiotemporal context (e.g., Transformer-based approaches) and the incorporation of explicit physical constraints, including energy and water conservation. Such discussion would help clarify how aiLand could evolve toward a more physically consistent and scalable land-surface emulator.

We thank the reviewer for this suggestion and have added Sect. 4.3, "Architecture and physical constraints", tied to results already in the paper: the resolution transfer of Appendix B, the architecture comparison of Wesselkamp et al. (2025), and the closure results of Sect. 3.2.2. One point the manuscript had left implicit is that ecLand is itself column-based, so the point-wise architecture matches the structure of the model being emulated; the limits of that argument, at higher resolution and under coupling to hydrological routing, are stated alongside it.

Added as Sect. 4.3:

4.3 Architecture and physical constraints

The applications above also bear on how the architecture itself should develop. The multi-layer perceptron used here operates point-wise and represents neither spatial context nor memory beyond the previous state. The absence of spatial context mirrors the structure of the target: ecLand solves each grid box as an independent column, with no lateral exchange within the land scheme, so a point-wise emulator matches the physical model rather than approximating a spatial one. In coupled mode, neighbouring columns interact only through the atmosphere. Because the model never sees the grid it is also largely resolution-agnostic (Appendix C), whereas transformer or graph-based architectures over the model grid would add representational capacity that the target does not use and would tie the learned

representation to a particular discretisation. This argument is bounded by resolution and by scope: lateral transport becomes increasingly relevant as grid spacing decreases, and the runoff and routing extension of Sect. 4.1 is where the land surface stops being column-independent, so an emulator coupled to a routing model would need to represent connectivity between grid boxes. Temporal context involves a different trade-off, since recurrence adds internal state that must be carried through the rollout. Wesselkamp et al. (2025) found a tuned LSTM advantageous for snow and upper-layer soil temperature over Europe, but this did not carry over to the global scale, where the MLP performed best of the three architectures tested and is the only one that is both differentiable and inexpensive to evaluate, as the applications in Sect. 4.2 require.

A second question is whether the emulator conserves energy and water. aiLand does not enforce conservation through loss penalties or architectural constraints, but neither is it unconstrained: it is trained on a model whose surface budgets close by construction, so approximate conservation is inherited as a learned property of the mapping. The results suggest this inheritance is substantial; the state remains coherent over four-year integrations (Sect. 3.1.1), and the fine-tuned variants improve energy balance closure although no closure term appears in the loss (Sect. 3.2.2). What learned consistency cannot offer is a guarantee, and it is most likely to fail in rare states, unseen regimes and long coupled integrations, where conservation matters most. Making the constraint explicit is not straightforward: the observational fine-tuning target is an energy-balance-closed product in which the residual is distributed by convention rather than measured (Sect. 2.1.4), so a model constrained to satisfy that closure would inherit the convention as a structural commitment. Whether hard constraints or soft penalties are preferable depends on the application, and we regard this as an open question for emulators of this kind.

2. Section 2.3.2: The rationale for introducing a new data-driven “biome” classification should be clarified. Since the clusters are defined using climate and land-surface variables, including soil moisture and Bowen ratio, they may be better described as environmental or climate–land-surface regimes rather than ecological biomes. I also encourage the authors to discuss why established classifications such as PFTs or Köppen–Geiger classes are insufficient and, if possible, examine whether the main transferability conclusions are robust to the choice of classification.

We agree on both points. The clusters are now called climate–land-surface regimes (R0–R5) throughout, and Sect. 2.3.2 opens with the rationale:

To test the transferability of aiLand-base, we partition the land surface into climate–land-surface regimes by clustering in a feature space that includes land-surface states (soil moisture, snow cover, Bowen ratio) as well as climate. Working in a metric feature space lets us relate transfer skill to the distance between regimes, and test whether a single global model matches regime specialists, both on regimes it has seen and on one withheld from training.

Established classifications are less suited to this analysis for two reasons. They do not come with a distance between classes, which the transfer analysis relies on (Fig. 7a). They are also defined on climate (Köppen–Geiger) or vegetation (PFTs) alone, whereas the emulator's errors are governed by the land-surface state: soil-moisture limitation and snow insulation are the two error sources identified in Sect. 3.1.4. PFTs are moreover a model construct, prescribed in ecLand rather than emerging from it, whereas our clustering uses realised state variables (i.e., soil moisture, snow cover and the Bowen ratio), which reflect how a grid box actually behaves, including where that departs from what its vegetation class would suggest.

To compare with an established classification, we have cross-tabulated the regimes against aggregated Köppen–Geiger groups (in a new Appendix C). The regimes are close to an aggregated Köppen classification. They split the tropical group into rainforest (R5) and seasonal humid (R0) regimes and the arid group into desert (R4) and steppe (R1), and they absorb most temperate C cells into R0 and polar E cells into R3.

We have not repeated the transfer experiments with a Köppen-based partition. It would differ from ours mainly by separating the temperate and polar classes, which the cross-tabulation shows behave like their neighbours, and by merging desert with steppe, which would blur the contrast behind the largest penalties in Fig. 7a. We therefore expect the main conclusions to hold. The leave-one-regime-out experiment (comment 3) supports this: the cost of omitting a regime depends on whether the remaining regimes contain similar conditions, not on where the boundaries are drawn.

The following table has been added to Appendix

Table A1. Composition of each regime by aggregated Köppen–Geiger group, as a percentage of the regime's grid cells (row-normalised). Groups contributing less than 6 % of a regime are omitted, so rows do not sum to 100 %.

Regime	Dominant Köppen–Geiger groups (% of regime cells)
R0 Tropical/subtropical humid	Tropical 41.7, Temperate-oceanic 17.6, Arid-hot 12.2, Temperate-monsoon 11.4
R1 Arid/semi-arid warm	Arid-hot 42.2, Arid-cold 28.8, Tropical 12.0
R2 Temperate continental	Continental-warm 57.7, Continental-cold 37.0
R3 Boreal/cold continental	Continental-cold 59.9, Polar 19.2, Arid-cold 15.5
R4 Dry desert	Arid-hot 91.4, Arid-cold 8.5
R5 Tropical rainforest	Tropical 81.5, Temperate-oceanic 10.0, Temperate-monsoon 6.3

3. The cross-biome transferability analysis may not provide a fully equivalent comparison because the biome-specialist models are evaluated on unseen biomes, whereas the global model has already been trained on all biomes, including the target biome. Thus, its better performance may partly reflect direct exposure to the target regime rather than stronger generalization. If computationally feasible, a leave-one-biome-out experiment, in which the global model is trained on all but the target biome, would provide a more rigorous test of transferability to unseen environmental regimes.

We thank the reviewer for this suggestion, which Reviewer 1 also raised. We have run the leave-one-regime-out experiment: six additional models, each trained on all regimes but one and evaluated on the held-out regime. Removing a regime degrades performance on it by at most 17 %, against specialist transfer penalties of up to 4.9×, so the global model's

advantage is not explained by exposure alone. The results are added as Table 7 and a paragraph in Sect. 3.1.3:

Table 6. Leave-one-regime-out (LORO) performance. Each LORO model is trained on all regimes but the target and evaluated on the held-out regime. RMSE is normalised by the mean in-regime specialist RMSE; global and in-regime specialist values are reproduced from Fig. 7a. Single training runs, so differences of a few per cent are within run-to-run variations.

Regime	LORO	Global	Specialist	LORO/global
R0 Tropical/subtropical humid	1.036	1.032	1.053	1.00
R1 Arid/semi-arid warm	1.565	1.467	1.140	1.07
R2 Temperate continental	0.967	0.976	1.089	0.99
R3 Boreal/cold continental	1.131	1.091	1.109	1.04
R4 Dry desert	1.844	1.596	0.707	1.16
R5 Tropical rainforest	0.838	0.799	0.903	1.05

Because the global model has seen every regime in training, its advantage over cross-regime specialist transfer could reflect exposure rather than generalisation. To separate the two, six leave-one-regime-out models were trained on all regimes but one and evaluated on the held-out regime (Table 6). Removing a regime degrades performance on it only slightly: four of six regimes are within 5 % of the global model, and the largest degradation is 16 % for R4, against cross-regime specialist penalties reaching 4.9×. The cost of omitting a regime does not scale with its share of the training data: removing R0, the largest regime in Table 3, has no effect, whereas removing R4, one of the smallest, is the most damaging. What matters is whether the remaining regimes contain the same physical conditions; R1 and R4 are the only water-limited regimes in the partition, so removing either removes most of the exposure to soil-moisture-limited evaporation.

We have revised the closing sentence of Sect. 3.1.3 accordingly:

Overall, a single global model provides the most robust performance across regimes, and this is not simply because it has seen every regime in training: removing a regime degrades performance on it by at most 16 %, against transfer penalties of up to 4.9× when a specialist is applied outside its training regime. What the global model gains is exposure to conditions (e.g., water limitation, snow-insulated soil) that no single-regime training set contains. Specialists offer an advantage primarily for diagnostic variables within climatically well-defined regimes.

and the Conclusions now state that the global model "loses at most 16 % of its skill on a regime withheld entirely from training".

4. Line 253: Please specify which validity metrics were used to determine the optimal number of clusters ($k = 7$).

The metrics are the silhouette score, the Davies–Bouldin index and the Calinski–Harabasz index. We have named them, with references, in Sect. 2.3.2:

Regimes were identified using k-means clustering (Lloyd, 1982). The number of clusters was chosen by evaluating three cluster validity metrics over $k = 2\text{--}12$: the silhouette score (Rousseeuw, 1987), the Davies–Bouldin index (Davies and Bouldin, 1979) and the Calinski–Harabasz index (Caliński and Harabasz, 1974). The first two are optimised at $k = 7$, which lies within 1.3 % of the optimum of the third at $k = 9$; we therefore use $k = 7$.

We have also added the following references to support this sentence:

Caliński, T. and Harabasz, J.: A dendrite method for cluster analysis, *Communications in Statistics*, 3, 1–27, 1974.

Davies, D. L. and Bouldin, D. W.: A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, 224–227, 1979.

Lloyd, S.: Least squares quantization in PCM, *IEEE Transactions on Information Theory*, 28, 129–137, 1982.

Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, 20, 53–65, 1987.