

Reviewer 1

This work introduces aiLand, an MLP-based emulator of ECMWF's land surface scheme ecLand. It rigorously evaluates its simulation performance and stability, then demonstrates its utility in two major experiments. The first is a cross-biome extrapolation exercise with unsupervised cluster detection. The second is a two-stage training framework that ablates finetuning approaches, transferring the physics-trained model to Eddy Covariance flux measurements, and evaluates for its skill to correct physical biases. Multiple sub-analyses are conducted, including energy balance closure at station sites.

Thanks, wonderful work and exciting to read! The manuscript is highly polished and in principle ready for publication. Given I was asked for a minor revisions review, I'd ask the authors to address the specific comments below and leave the general comments to their consideration.

Thank you for the careful and constructive reading, and for the positive assessment. We are glad that the work was enjoyable to read.

General comments

An operational perspective drives the experimental design with two cool major experiments. Yet, one of them gets buried in the introduction and together they remain slightly fragmented. The broader context that is opened in the introduction remains underdiscussed, restricted to a short sentence in the conclusion. Better integrating them in the discussion could — aside releasing this new model version — make a strong case for pre-training on physical models when forecasting land states under (eco)climatic regime changes.

This is very good point, and we have added two sentences opening Sect. 4 that tie the two experiments together before the perspectives that follow. We have not added a discussion section, since the manuscript is already quite long.

Added at the start of Sect. 4, before Sect. 4.1:

Read together, the cross-regime and fine-tuning experiments make a case for physical pretraining that goes beyond the emulation of one model: the mapping learned from ecLand transfers across regimes and survives the removal of a regime from the training data (Table 6). In addition, it is retained rather than overwritten when adjusted towards sparse observations. Transferable physical structure plus observational correction is what an ML land surface component would require where the observational record is sparse.

The finetuned emulator is evaluated for its bias correction, which will be useful as a coupled model component in NWP or AIWP. The introduction and experimental setup (especially experiment 1) also touch on the field of studying transfer learning capability of physically (weak-) constrained ML. The pretrained aiLand together with experiment 1 is in principle a perfect candidate to demonstrate if weak physical priors allow ML

prediction under covariate shifts — with implications for upscaled products. But the physics pre-trained model would need to be compared to an (afap) unconstrained baseline, the closest to which I believe is your scenario S5. As you also deem S5 to yield the best return in generalization skill, I would have liked to see this topic at least discussed, starting from what you'd expect if you trained aiLand on Flux stations directly. You use stations with >20 years of observations, an amount that may be enough for an MLP to learn better than an emulation of a biased target when physical model parameters are constant in time. Perhaps generalization or drift at seasonal lead times if done prognostically will become a problem, and if a learned physical prior reduces it, that would make a strong case for establishing the two-stage training regime.

None of the fine-tuning strategies (Table 2) are weakly constrained, S5 included: FLUXNET replaces only the two flux targets, and all other outputs remain supervised by ecLand (Sect. 2.1.5). It is possible to train an empirical model directly on the stations data, and models of this kind (i.e., trained on tower forcing to predict LE and H) are the empirical benchmarks used in PLUMBER2 (Abramowitz et al., 2024), where they are compared against land surface models at the same sites. Such a model has no prognostic state, however, so it cannot be assessed for drift (Sect. 3.2.3) and answers a different question from the one posed here. We did train the closest variant that keeps aiLand's structure, the same architecture trained from random initialisation on the fine-tuning dataset alone. At the tower sites its latent heat flux error is lower than ecLand's, but its LE and H biases are large and of opposite sign, so they partly cancel in the total turbulent flux. The prognostic state is where it fails: soil moisture departs from ecLand three to six times more than aiLand-base does, and the deep soil dries over the evaluation period while ecLand's does not. This is consistent with the reviewer's suggestion that the learned prior limits drift, although a direct test would need observational constraints on the state itself, which we identify as future work in the Conclusions. We have added to Sect. 3.2.4:

The same architecture trained from scratch on the fine-tuning dataset improves latent heat flux at the tower sites, but with opposing LE and H biases and a soil-moisture departure from ecLand three to six times that of aiLand-base, suggesting that the prognostic stability of aiLand-base comes from pretraining on the global grid.

The cross-biome experiment gets buried a bit across all sections in the manuscript, and most importantly, its goal and implications are not really stated. The methods and results section lack clarity and may profit from a formalization of the analysis that is shown in the results. My assumption about the goal is demonstrating the advantage of a global model, as the experiment is inversely conducted to a spatial cross validation — which in contrast would give an estimate of the true extrapolation error but means training on all but one biome and predict on the hold-out (i.e. Roberts et al. 2016). The approach to manually identify clusters from features that are derived from the training data is a strength of this experiment: it could provide controlled extrapolation regimes, but we don't see where exactly the clusters are subsets of the global covariate distribution and where not, hence where biome-specialists have to extrapolate. If this was not the goal, it raises the question, why not use established biome classifications (and their naming), or at least validate detected clusters with such. The identified

clusters are meaningful, but their labels blur these advantages and instead highlight limitations.

We agree that the goal was not stated. It is twofold: i) to relate transfer skill to the dissimilarity between training and evaluation regimes, which needs a partition in a metric feature space (the x-axis of Fig. 7a); and ii) to test whether a single global model matches regime specialists. Established classifications do not come with such a distance and are defined on climate or vegetation alone, whereas the emulator's difficulty depends on the land-surface state, which is why soil moisture, snow cover and Bowen ratio are among the clustering features.

We have stated this at the start of Sect. 2.3.2 (below). We have also run the leave-one-regime-out experiment suggested by this reviewer (and reviewer 2) and reported it via a new table in Section 3.1.3. Percentile ranges were added to Table 3 to show where regimes overlap. Finally, we have cross-tabulated the regimes against aggregated Köppen–Geiger groups (see in the Appendix). The regimes are close to an aggregated Köppen classification, with the tropical group split into rainforest and seasonal humid regimes, the arid group split into desert and steppe, and the temperate and polar groups absorbed into R0 and R3.

Replacing the first sentence of Sect. 2.3.2

To test the transferability of aiLand-base, we partitioned the land surface into climate–land-surface regimes by clustering those in a feature space that combines climatological indicators with land-surface states (soil moisture, snow cover, Bowen ratio). 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.

Specific comments

L 362: Is this asymmetry a modelling or a data result? Given that you target two of the diagnostics for finetuning on observational data, the autoregressive error structure of which will be real. I see that the diagnostic prediction is an advantage for gappy data prediction, perhaps avoids unstable rollouts — but was it ever considered to model your diagnostics prognostically, as does the AIFS with e.g. skin temperature?

The asymmetry is a modelling result. Diagnostics are predicted as absolute values from the current state and forcing at each step, with no memory of their own previous value, so they cannot accumulate errors along the rollout whichever target they are trained on. Their RMSE is accordingly almost constant with lead time (Table 4), whereas that of the prognostic states grows.

We did consider a prognostic treatment. Beyond the advantages the reviewer notes for gappy data and rollout stability, we kept the diagnostic formulation because it mirrors ecLand, in which skin temperature and the turbulent fluxes are computed from the tile energy balance at each time step rather than carried as prognostic states; treating them prognostically would add a state with no counterpart in the model being emulated. In the AIFS, by contrast, skin temperature is one of the surface fields predicted and fed

back as part of the jointly learned model state. We have added one sentence to Sect. 2.2.1 (after l. 175):

Diagnostic variables are predicted as absolute values without memory of their previous value, mirroring ecLand, where skin temperature and the turbulent fluxes are computed from the tile energy balance at each time step rather than carried over from the previous one as prognostic states.

Section 2.3.2: Naming conventions are useful for interpretation, but because they are established in existing biome and climate classifications, they are also confusing. E.g. the assignment of large parts of Europe to a subtropical humid cluster, instead of the temperate one. I don't have a better suggestion though but strictly using their numbered labels throughout.

We have followed the reviewer's suggestion and the regimes are now referred to as R0–R5 throughout the text, figure legends and tables. The descriptive names remain only in Table 3 and the Fig. 2 legend, with the following caveat added to the Table 3 caption:

Descriptive names are interpretive labels for the dominant character of each cluster and do not correspond to established biome or climate classes (overlaps with Köppen–Geiger group are given in Table C1).

L 250: Why was the period of the training feature space only partially considered for feature computation?

The 2009–2020 window was a practical choice. Since the features are climatological, the partition is insensitive to it: recomputing them over the full training period (2000–2019) or a disjoint earlier decade (2000–2008) assigns 98.9 % and 97.6 % of land points, respectively, to the same regime. We have added this to Sect. 2.3.2:

The features define static spatial regions, so a decade-long window suffices; computing them over the full training period instead assigns 98.9 % of land points to the same regime, which does not provide a justifiable improvement over the shorter period.

L 253: please specify the validity metrics.

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.

Table 3: Reporting e.g. maximum and minimum could better clarify where the models are extrapolating in the cross-biome exercises.

This is indeed a good suggestion. We have added the 5th–95th percentile range of mean annual temperature and precipitation to Table 3. We decided to report percentiles rather than minima and maxima because the extremes are set by a handful of non-representative coastal and orographic cells. The percentile ranges are computed on the grid-point distribution of climatological means within each cluster, from the same variables and the same $k = 7$ clustering that produced the published means, so the mean columns remain unchanged.

Revised Table 3:

Regime	Primary Character	Temp (°C)	Temp 5–95 %	Precip (mm)	Precip 5–95 %	LAI	Snow (%)	Points (%)
R0	Tropical/Subtropical Humid	21	10–28	1050	430–1750	4.3	1%	27.7%
R1	Arid/Semi-Arid Warm	18	0–28	640	120–2330	1.2	3%	20.7%
R2	Temperate Continental	3	–5–11	780	410–1350	3.9	37%	14.9%
R3	Boreal/Cold Continental	–4	–12–7	550	260–1060	2.7	52%	13.4%
R4	Dry Desert	24	13–30	58	3–185	0.4	0%	11.9%
R5	Tropical Rainforest	24	14–27	2700	1550–4600	6.1	1%	11.4%
R6	Glaciers (masked)	–	–	–	–	–	–	–

Added to the Table 3 caption:

The provided ranges are the 5th–95th percentiles of the grid-point distribution of climatological means within each regime.

Added to Sect. 3.1.3 (with the discussion of which regimes are hardest):

R4 emerges as the most challenging target overall (mean 3.1x **consistent with its outlier climatology (Table 3): on the precipitation axis it overlaps only with R1, so transfers into it are extrapolations for every other specialist, whereas R1, spanning 120-2330 mm, overlaps both the desert and the humid regimes.**

L 275 / section 3.1.3 / L 628, ...: I suggest using consistent names throughout the manuscript — replace ecosystem regimes with biomes, also because it's climate-defined clusters.

We agree on consistency. Since Reviewer 2 raised the same point but recommended “regimes” over “biomes” on the grounds that the clusters are defined on climate and land-surface state rather than ecology, we have standardised on “regimes” throughout (“climate–land-surface regimes” at first mention; “ecosystem regimes”, “biomes” and “ecosystem robustness” are removed). Section 3.1.3 is retitled “Cross-regime robustness”.

L 284: Is this characterizing short-range or seasonal error growth?

We thank the reviewer for pointing this out – the 90-day forecast protocol is indeed used to characterise short- to sub-seasonal error growth, we made this explicit in the text:

...used to characterise short-~~range~~ **to sub-seasonal** error growth

L 285: [...] diverge under identical forcing but different initial conditions.

In our experiments aiLand-base and ecLand are initialised from the same state and driven by identical forcing, so the error growth reflects the emulator departing from ecLand rather than sensitivity to initial conditions. We have clarified this at l. 250:

as aiLand-base ~~and ecLand diverge under~~ **departs from ecLand when both are initialised from the same state and driven by** identical forcing.

Table 4: Do you know why you accumulate more error in deeper layers, where the relative error is smaller?

Errors accumulate at depth because the deep layers are weakly coupled to the atmospheric forcing: at the surface. Errors in the state are continually corrected by the strong dependence of the next increment on the forcing, whereas deep-layer increments depend mostly on the state itself. The large growth ratios are further amplified by the very small day-1 errors at depth, as already noted at l. 253–255. We have added one sentence after l. 255:

Unlike the surface layer, where errors are continually corrected by the forcing, deep-layer increments depend mostly on the state itself, so errors there accumulate rather than being corrected.

L 308: Interesting. Is this skewness only showing for stl1?

No, skewness is also shown for the other layers, but stl1 is the only layer where the skew is substantial and robust (0.94, ≈ 0.9 after excluding the most extreme 1 % of grid points). stl2 is less skewed (0.42). For stl3, the skewness computed over all points is ≈ 3.6 , but this comes entirely from a few extreme increments at coastal and mountain points; excluding the most extreme 1 % brings it to ≈ 0 . The bias is correspondingly largest in stl1 (+0.40 K, against +0.14 and +0.17 K below). We have added one sentence at l.:

The skew weakens with depth (0.42 for stl2, and approximately zero for stl3 once the most extreme 1% of grid points is excluded), consistent with the bias being largest in the surface layer, which receives the diurnal forcing directly.

L 309: Should skewness not be unitless, or did you compute this as the difference between mean and median? Then this is large for increments.

The reviewer is right pointing out that skewness is dimensionless; the "K" was an editorial mistake and has been removed. The value is the standard Fisher–Pearson coefficient, $g_1 = m_3/m_2^{3/2}$, not a mean-minus-median difference. Read as such, 0.94 is a moderate right skew rather than a large one, plausibly reflecting faster daytime warming than night-time cooling in the surface layer.

Figure 7a: Ngl, it takes a while to follow the paragraph that describes these results. Part of it is terminology that is not uniquely defined (i.e. generalization error = transfer skill), or used inconsistently (L 354: biome naming conventions already mentioned, tropical vs. tropical/subtropical, L 359 tropical or rainforest — other paragraphs just use the shortages, maybe preferable).

We have defined the quantities once in Sect. 2.3.2, replaced "generalisation error", "transfer skill" and "cross-biome degradation" with a single term, "transfer penalty", and used the R0–R5 labels throughout Sects. 3.1.3 and 3.2 and the Fig. 7 caption.

Figure 7b: I don't understand 7b from the caption description alone: What is the residual penalty under global training in contrast to the penalty recovered by the global model? How are the in- and cross biome evaluations related with the global model? Could you please define and explain these conventions somewhere, e.g. in section 2.3.2?

We agree that the definitions were missing. Every model, including the global one, is evaluated on every regime, and each regime's own specialist provides the baseline. For each variable, the full bar in Fig. 7b is the mean excess error of specialists applied outside their training regime; the red portion is the excess error the global model still incurs on the same regimes, and the blue portion the error avoided by training on all regimes. We have defined the transfer penalty at the end of Sect. 2.3.2:

Each specialist and the global model is evaluated on every regime, with RMSE normalised per variable by the mean in-regime specialist RMSE, so that 1.0 corresponds to average specialist skill. The transfer penalty of a model on a regime is its normalised RMSE there relative to that of the regime's own specialist, minus one.

and rewritten the Fig. 7b caption to explain the decomposition:

(b) Mean transfer penalty per variable (Sect. 2.3.2) of specialists applied outside their training regime. The red portion is the penalty the global model still incurs on the same regimes (residual penalty); the blue portion is the difference (recovered by global training). Numbers give the total cross-regime RMSE ratio (penalty + 1).

L 360: It could be helpful to visualize the statements in the plot, i.e. highlight the (tropical) rainforest-to-desert pixel and desert cluster.

Done. The R5→R4 point is annotated in Fig. 7a (right), and the points evaluated on R4 are outlined so that the desert-as-target cluster is visible.

L 368: This paragraph is most supported by figure 7a left panel and would fit after the one describing the 7a right panel results.

Agreed; the paragraph has been moved accordingly.

Table 6: I think an aiLand-S0 (Flux) is missing from this table, can be discussed.

We take S0 (Flux) to mean a model trained on flux observations without pretraining. An empirical mapping forcing and static fields directly to LE and H is possible, but it is the observation-only approach already benchmarked in PLUMBER2 (Abramowitz et al., 2024): it has no prognostic state and so addresses a different question from aiLand. The closest variant that keeps aiLand's structure, the same architecture trained from random initialisation on the fine-tuning dataset, is still trained against ecLand for all outputs other than LE and H, so it does not remove the physical constraint; we discuss it briefly in Sect. 3.2.4 rather than in Table 6.